Density Forecasts in Panel Data Models:

A Semiparametric Bayesian Perspective*

Laura Liu†
University of Pennsylvania

April 28, 2017

Abstract

This paper constructs individual-specific density forecasts for a panel of firms or households using a dynamic linear model with common and heterogeneous coefficients and cross-sectional heteroskedasticity. The panel considered in this paper features large cross-sectional dimension (N) but short time series (T). Due to short T, traditional methods have difficulty in disentangling the heterogeneous parameters from the shocks, which contaminates the estimates of the heterogeneous parameters. To tackle this problem, I assume that there is an underlying distribution of heterogeneous parameters, model this distribution nonparametrically allowing for correlation between heterogeneous parameters and initial conditions as well as individual-specific regressors, and then estimate this distribution by pooling the information from the whole cross-section together. I develop a simulation-based posterior sampling algorithm specifically addressing the nonparametric density estimation of unobserved heterogeneous parameters. I prove that both the estimated common parameters and the estimated distribution of the heterogeneous parameters achieve posterior consistency, and that the density forecasts asymptotically converge to the oracle forecast, an (infeasible) benchmark that is defined as the individual-specific posterior predictive distribution under the assumption that the common parameters and the distribution of the heterogeneous parameters are known. Monte Carlo simulations demonstrate improvements in density forecasts relative to alternative approaches. An application to young firm dynamics also shows that the proposed predictor provides more accurate density predictions.

JEL Codes: C11, C14, C23, C53, L25

Keywords: Bayesian, Semiparametric Methods, Panel Data, Density Forecasts, Posterior Consistency, Young Firms Dynamics

---

*First version: November 15, 2016. Latest version: https://goo.gl/c6Ybrd. I am indebted to my advisors, Francis X. Diebold and Frank Schorfheide, for much help and guidance at all stages of this research project. I also thank the other members of my committee, Xu Cheng and Francis J. DiTraglia, for their advice and support. I further benefited from many helpful discussions with Gianni G. Amisano, Evan Chan, Timothy Christensen, Benjamin Connault, Jeremy Greenwood, Hyungsik R. Moon, Andriy Norets, Paul Sangrey, Minchul Shin, Molin Zhong, and seminar participants at the University of Pennsylvania, the Federal Reserve Bank of Philadelphia, the Federal Reserve Board, the University of Virginia, Microsoft, the University of California, Berkeley, the University of California, San Diego (Rady), Boston University, and the University of Illinois at Urbana–Champaign, as well as conference participants at the 26th Annual Meeting of the Midwest Econometrics Group. I would also like to acknowledge the Kauffman Foundation and the NORC Data Enclave for providing researcher support and access to the confidential microdata. All remaining errors are my own.

†Department of Economics, University of Pennsylvania, Philadelphia, PA 19104. Email: yuliu4@sas.upenn.edu.

1 Introduction

Panel data, such as a collection of firms or households observed repeatedly for a number of periods, are widely used in empirical studies and can be useful for forecasting individuals' future outcomes, which is interesting and important in many applications. For example, PSID can be used to analyze income dynamics (Hirano, 2002; Gu and Koenker, 2015), and bank balance sheet data can help conduct bank stress tests (Liu *et al.*, 2016). This paper constructs individual-specific density forecasts using a dynamic linear panel data model with common and heterogeneous parameters and cross-sectional heteroskedasticity.

In this paper, I consider young firm dynamics as the empirical application. For illustrative purposes, let us consider a simple dynamic panel data model as the baseline setup:

$$\underbrace{y_{it}}_{\text{performance}} = \beta y_{i,t-1} + \underbrace{\lambda_i}_{\text{skill}} + \underbrace{u_{it}}_{\text{shock}}, \quad u_{it} \sim N\left(0, \sigma^2\right), \tag{1.1}$$

where $i = 1, \cdots, N$, and $t = 1, \cdots, T+1$. The y_{it} is the observed firm performance such as the log of employment,[1] λ_i is the unobserved skill of an individual firm, and u_{it} is an i.i.d. shock. Skill is independent of the shock, and the shock is independent across firms and times. β and σ^2 are common across firms, where β represents the persistence of the dynamic pattern, and σ^2 gives the size of the shocks. Based on the observed panel from period 0 to period T, I am interested in forecasting the future performance of any specific firm in period $T+1$, $y_{i,T+1}$.

The panel considered in this paper features large cross-sectional dimension N but short time series T. This framework is appealing to the young firm dynamics example because the number of observations for each young firm is restricted by its age. Good estimates of the unobserved skill λ_is facilitate good forecasts of $y_{i,T+1}$s. Due to short T, traditional methods have difficulty in disentangling the unobserved skill λ_i from the shock u_{it}, which contaminates the estimates of λ_i. The naive estimators that only utilize the firm-specific observations are inconsistent even if N goes to infinity.

To tackle this problem, I assume that λ_i is drawn from the underlying skill distribution f and estimate this distribution by pooling the information from the whole cross-section together. In terms of modeling f, the parametric Gaussian density misses many features in real world data, such as asymmetricity, heavy tails, and multiple peaks. For example, since good ideas are scarce, the skill distribution of young firms may be highly skewed. In this sense, the challenge now is how we can model f more carefully and flexibly. Here I estimate f via a nonparametric Bayesian approach where the prior is constructed from a mixture model and allows for correlation between λ_i and the initial condition y_{i0} (i.e. a correlated random effects model).

Once this distribution is estimated, I can, intuitively speaking, use it as a prior distribution and

[1] Employment is one of the standard measures in the firm dynamics literature (Akcigit and Kerr, 2010; Zarutskie and Yang, 2015).

update it with the firm-specific data and obtain the firm-specific posterior. In a special case where the common parameters are set to be $\beta = 0$ and $\sigma^2 = 1$, the firm-specific posterior is characterized by Bayes' theorem,

$$p\left(\lambda_i \left| f, y_{i,0:T}\right.\right) = \frac{p\left(y_{i,1:T} \left| \lambda_i\right.\right) f\left(\lambda_i \left| y_{i0}\right.\right)}{\int p\left(y_{i,1:T} \left| \lambda_i\right.\right) f\left(\lambda_i \left| y_{i0}\right.\right) d\lambda_i}. \tag{1.2}$$

This firm-specific posterior helps provide a better inference about the unobserved skill λ_i of each individual firm and a better forecast of the firm-specific future performance, thanks to the underlying distribution f that integrates the information from the whole panel in an efficient and flexible way.[2]

It is natural to construct density forecasts based on the firm-specific posterior. In general, forecasting can be done in point, interval, or density fashion, whereas density forecasts give the richest insight regarding future outcomes. By definition, a density forecast provides a predictive distribution of firm i's future performance and summarizes all sources of uncertainties, hence is preferable in the context of young firm dynamics and other applications with large uncertainties and nonstandard distributions. In particular, for the dynamic panel data model as specified in equation (1.1), the density forecasts reflect uncertainties arising from future shock $u_{i,T+1}$, individual heterogeneity λ_i, and estimation uncertainty of common parameters $\left(\beta, \sigma^2\right)$ and skill distribution f.

A typical question that density forecasts could answer is: what is the chance that firm i will hire 5, 10, or 100 more people next year? The answer to this kind of question is valuable to both investors and regulators regarding how promising or troublesome each firm could be. For investors, it is helpful to select a better performing portfolio of startups.[3] For regulators, more accurate forecasts facilitate monitoring and regulation of bank-lending practices and entrepreneur funding.[4] Moreover, once the density forecasts are obtained, one can easily recover the point and interval forecasts.

A benchmark for evaluating density forecasts is the posterior predictive distribution for $y_{i,T+1}$ under the assumption that the common parameters $\left(\beta, \sigma^2\right)$ and the distribution of the heterogeneous coefficients f are known. I refer to this predictive density as the (infeasible) oracle forecast. In the special case where $\beta = 0$ and $\sigma^2 = 1$, it is straightforward to construct the oracle predictor for firm i, which combines firm i's uncertainties due to future shock and heterogeneous skill.

$$f_{i,T+1}^{oracle}\left(y\right) = \int \underbrace{\phi\left(y - \lambda_i\right)}_{\text{future shock}} \cdot \underbrace{p\left(\lambda_i \left| f_0, y_{i,0:T}\right.\right)}_{\text{heterogeneous skill}} \cdot d\lambda_i.$$

The part of skill uncertainty is exactly the firm-specific posterior in equation (1.2) and arises from the lack of time-series information available to infer individual λ_i. Therefore, the common skill

[2]Note that this is only an intuitive explanation why the skill distribution f is crucial. In the actual implementation, the estimation of the correlated random effect distribution f, the estimation of common parameters $\left(\beta, \sigma^2\right)$, and the inference of firm-specific skill λ_i are all done simultaneously.

[3]The general model studied can include aggregate variables that have heterogeneous effects on individual firms, so their coefficients can be thought of as the betas for portfolio choices.

[4]The aggregate-level forecasts can be obtained by summing firm-specific forecasts over different subgroups.

distribution f_0 helps in formulating firm i's skill uncertainty and contributes to firm i's density forecasts through the channel of skill uncertainty.

In practice, however, the skill distribution f (as well as the common parameters for models beyond the special case) is unknown and unobservable, thus introducing another source of uncertainty. Now the oracle predictor becomes an infeasible optimum. A good feasible predictor should be as close to the oracle as possible, which in turn calls for a good estimate of the underlying skill distribution f. The proposed semiparametric Bayesian procedure achieves better estimates of the underlying skill distribution f than parametric approaches, hence more accurate density forecasts of the future outcomes. In the special case where $\beta = 0$ and $\sigma^2 = 1$, the three sources of uncertainties can be decomposed as follows:[5]

$$f_{i,T+1}^{sp}(y) = \int \underbrace{\phi(y - \lambda_i)}_{\text{future shock}} \cdot \underbrace{p(\lambda_i \,|f, y_{i,0:T})}_{\text{heterogeneous skill}} \cdot \underbrace{d\Pi(f \,|y_{1:N,0:T})}_{\text{estimation}} d\lambda_i.$$

The contributions of this paper are threefold. First, I develop a posterior sampling algorithm specifically addressing nonparametric density estimation of the unobserved λ_i. For a random effects model, which is a special case with zero correlation between λ_i and y_{i0}, the f part becomes a relatively simple unconditional density estimation problem. I impose a Dirichlet Process Mixture (DPM) prior on f and construct a posterior sampler building on the blocked Gibbs sampler proposed by Ishwaran and James (2001, 2002). For a correlated random effects model, I further adapt the proposed algorithm to the much harder conditional density estimation problem using a probit stick breaking process prior suggested by Pati *et al.* (2013).

Second, I establish the theoretical properties of the proposed semiparametric Bayesian predictor when the cross-sectional dimension N tends to infinity. First, I provide conditions for identifying both the parametric component (β, σ^2) and the nonparametric component f. Second, I prove that both the estimated common parameters and the estimated distribution of the heterogeneous coefficients achieve posterior consistency, an essential building block for bounding the discrepancy between the proposed predictor and the oracle. Compared to previous literature on posterior consistency, there are several challenges in the current setting: (1) disentangling unobserved individual effects λ_is and shocks u_{it}s, (2) incorporating an unknown shock size σ^2, (3) adding lagged dependent variables as covariates, and (4) addressing correlated random effects from a conditional density estimation point of view. Finally, I show that the density forecasts asymptotically converge to the oracle forecast in weak topology, which constitutes another contribution to the nonparametric Bayesian literature and specifically designed for density forecasts.

To accommodate many important features of real-world empirical studies, I extend the simple model (1.1) to a more general specification. First, a realistic application also incorporates other observables with common effects $(\beta' x_{i,t-1})$, where $x_{i,t-1}$ can include lagged y_{it}. Second, it is helpful to

[5]The superscript "sp" stands for "semiparametric".

consider observables with heterogeneous effects ($\lambda_i' w_{i,t-1}$), i.e. a correlated random coefficients model. Finally, beyond heterogeneity in coefficients (λ_i), it is desirable to take into account heterogeneity in shock sizes (σ_i^2) as well.[6] All numerical methods and theoretical properties are further established for the general specification.

Third, Monte Carlo simulations demonstrate improvements in density forecasts relative to predictors with various parametric priors on f, evaluated by log predictive score. An application to young firm dynamics also shows that the proposed predictor provides more accurate density predictions. The better forecasting performance is largely due to three key features (in order of importance): the nonparametric Bayesian prior, cross-sectional heteroskedasticity, and correlated random coefficients. The estimated model also helps shed light on the latent heterogeneity structure of firm-specific coefficients and cross-sectional heteroskedasticity, as well as whether and how these unobserved heterogeneous features depend on the initial condition of the firms.

It is worth mentioning that although I describe the econometric intuition using the young firm dynamics application as an example, the method can be applied to many economic and financial analyses that feature panel data with relatively large N and small T, such as microeconomic panel surveys (e.g. PSID, NLSY, and Consumer Expenditure Survey (CE)), macroeconomic sectoral and regional panel data (e.g. Industrial Production (IP), and State and Metro Area Employment, Hours, and Earnings (SAE)), and financial institution performance (e.g. Commercial Bank Data and Holding Company Data). Which T can be considered as a small T depends on the dimension of individual heterogeneity (d_w), the cross-sectional dimension (N), and size of the shocks (σ^2 or σ_i^2). There can still be a significant gain in density forecasts even when T exceeds 100. Roughly speaking, the proposed predictor would provide sizeable improvement as long as the time series for individual i is not informative enough to fully reveal its individual effects, λ_i and σ_i^2.

Moreover, the method proposed in this paper is general to many other problems beyond forecasting. Here estimating heterogeneous parameters is important because we want to generate good forecasts, but in other cases, the heterogeneous parameters themselves can possibly be the objects of interest. For example, people may be interested in individual-specific treatment effects, and the technique developed here can be applied to those questions.

Related Literature First, this paper contributes to the literature on individual forecast in a panel data setup, and is closely related to Liu *et al.* (2016) and Gu and Koenker (2015, 2016). Liu *et al.* (2016) focus on point forecasts. They utilize the idea of Tweedie's formula to steer away from the complicated deconvolution problem in estimating λ_i. Unfortunately, the Tweedie shortcut is not applicable to the inference of underlying λ_i distribution and therefore not suitable for density forecasts.

[6] Here and below, the terminologies "random effects model" and "correlated random effects model" also apply to individual effects on σ_i^2, which are slightly different from the traditional definitions concentrated on λ_i.

Gu and Koenker (2015) address the density estimation problem. Their method is different from the one proposed in this paper in that this paper infers the underlying λ_i distribution via a full Bayesian approach (i.e. imposing a prior on the λ_i distribution and updating the prior belief by the observed data), whereas they employ an empirical Bayes procedure (i.e. picking the λ_i distribution by maximizing the marginal likelihood of data). In principle, the full Bayesian approach is preferable for density forecasts as it captures all kinds of uncertainties, including estimation uncertainty of the underlying λ_i distribution, which has been omitted by the empirical Bayes procedure. In addition, this paper features correlated random effects allowing for both cross-sectional heterogeneities and cross-sectional heteroskedasticities interacting with the initial conditions, whereas the Gu and Koenker (2015) approach focuses on random effects models without such interaction.

In their recent paper, Gu and Koenker (2016) also compare their method with an alternative nonparametric Bayesian estimator featuring a Dirichlet Process (DP) prior under a set of fixed scale parameters. There are two major differences between their DP setup and the DPM prior used in this paper. First, the DPM prior provides continuous individual effect distributions, which is more reasonable in many empirical setups. Second, unlike their set of fixed scale parameters, this paper incorporates a hyperprior for the scale parameter and updates it via the observed data, hence let the data choose the complexity of the mixture approximation, which can essentially be viewed as "automatic" model selection.[7]

There have also been empirical works on the DPM model with panel data, such as Hirano (2002), Burda and Harding (2013), Rossi (2014), and Jensen *et al.* (2015), but they focus on empirical studies rather than theoretical analysis. Hirano (2002) and Jensen *et al.* (2015) use linear panel models, while their setups are slightly different from this paper. Hirano (2002) considers flexibility in u_{it} distribution instead of λ_i distribution. Jensen *et al.* (2015) assume random effects instead of correlated random effects. Burda and Harding (2013) and Rossi (2014) implement nonlinear panel data models via either a probit model or a logit model, respectively.

Among others, Delaigle *et al.* (2008) have also studied the similar deconvolution problem and estimated the λ_i distribution in a frequentist way, but the frequentist approach misses estimation uncertainty, which matters in density forecasts, as mentioned previously.

Second, in terms of asymptotic properties, this paper relates to the literature on posterior consistency of nonparametric Bayesian methods in density estimation problems. The pioneer work by Schwartz (1965) lays out two high-level sufficient conditions in a general density estimation context. Ghosal *et al.* (1999) bring Schwartz (1965)'s idea into the analysis of density estimation with DPM priors. Amewou-Atisso *et al.* (2003) extend the discussion to linear regression problems with an unknown error distribution. Tokdar (2006) further generalizes the results to cases in which the true density has heavy tails. For a more thorough review and discussion on posterior consistency in Bayesian nonparametric problems, please refer to the handbooks, Ghosh and Ramamoorthi (2003)

[7]Section 6 shows the simulation results comparing the DP prior vs the DPM prior, where both incorporate a hyperprior for the scale parameter.

and Hjort *et al.* (2010) (especially Chapters 1 and 2). To handle conditional density estimation, similar mixture structure can be implemented, where the mixing probabilities can be characterized by a multinomial choice model (Norets, 2010; Norets and Pelenis, 2012), a kernel stick break process (Norets and Pelenis, 2014; Pelenis, 2014), or a probit stick breaking process (Pati *et al.*, 2013). I adopt the Pati *et al.* (2013) approach to offer a more coherent nonparametric framework that is totally flexible in the conditional measure. This paper builds on these previous works and establishes the posterior consistency result for panel data models. Furthermore, this paper obtains the convergence of the semiparametric Bayesian predictor to the oracle predictor, which is another new finding to the literature and specific to density forecasts.

Third, the algorithms constructed in this paper build on the literature on the posterior sampling schemes for DPM models. The vast Markov chain Monte Carlo (MCMC) algorithms can be divided into two general categories. One is the Pólya urn style samplers that marginalize over the unknown distribution G (Escobar and West, 1995; Neal, 2000).[8] The other resorts to the stick breaking process (Sethuraman, 1994) and directly incorporates G into the sampling procedure. This paper utilizes a sampler from the second category, Ishwaran and James (2001, 2002)'s blocked Gibbs sampler, as a building block for the proposed algorithm. Basically, it incorporates truncation approximation and augments the data with auxiliary component probabilities, which helps break down the complex posterior structure and thus enhance mixing properties as well as reduce computation time.[9] I further adapt the proposed algorithm to the conditional density estimation for correlated random effects using the probit stick breaking process prior suggested by Pati *et al.* (2013).

Last but not least, the empirical application in this paper also links to the young firm dynamics literature. Akcigit and Kerr (2010) document the fact that R&D intensive firms grow faster, and such boosting effects are more prominent for smaller firms. Robb and Seamans (2014) examine the role of R&D in capital structure and performance of young firms. Zarutskie and Yang (2015) present some empirical evidence that young firms experienced sizable setbacks during the recent recession, which may partly account for the slow and jobless recovery. For a thorough review on young firm innovation, please refer to the handbook by Hall and Rosenberg (2010). The empirical analysis of this paper builds on these previous findings. Besides providing more accurate density forecasts, we can also use the estimated model to analyze the latent heterogeneity structure of firm-specific coefficients and cross-sectional heteroskedasticity, as well as whether and how these unobserved heterogeneous features depend on the initial condition of the firms.

The rest of the paper is organized as follows. Section 2 introduces the baseline panel data model, the predictors for density forecasts, and the nonparametric Bayesian priors. Section 3 proposes the posterior sampling algorithms. Section 4 characterizes identification conditions and large sample properties. Section 5 presents various extensions of the baseline model together with correspond-

[8] For the definition of G, see equation (2.5).

[9] Robustness checks have been conducted with the more sophisticated slice-retrospective sampler (Dunson, 2009; Yau *et al.*, 2011; Hastie *et al.*, 2015), which does not involve hard truncation but is more complicated to implement. Results from the slice-retrospective sampler are comparable with the simpler truncation sampler.

ing algorithms and theorems. Section 6 examines the performance of the semiparametric Bayesian predictor using simulated data, and Section 7 applies the proposed predictor to the confidential microdata from the Kauffman Firm Survey and analyzes the empirical findings on young firm dynamics. Finally, Section 8 concludes and sketches future research directions. Notations, proofs, as well as additional algorithms and results can be found in the Appendix.

2 Model

2.1 Baseline Panel Data Model

The baseline dynamic panel data model is specified in equation (1.1),

$$y_{it} = \beta y_{i,t-1} + \lambda_i + u_{it}, \quad u_{it} \sim N\left(0, \sigma^2\right),$$

where $i = 1, \cdots, N$, and $t = 1, \cdots, T + h$. The y_{it} is the observed individual outcome, such as young firm performance. The main goal of this paper is to estimate the model using the sample from period 1 to period T and forecast the future distribution of $y_{i,T+h}$. In the remainder of the paper, I focus on the case where $h = 1$ (i.e. one-period-ahead forecasts) for notation simplicity, but the discussion can be extended to multi-period-ahead forecasts via either a direct or an iterated approach (Marcellino *et al.*, 2006).

In this baseline model, there are only three terms on the right hand side. $\beta y_{i,t-1}$ is the AR(1) term on lagged outcome, which captures the persistence pattern. λ_i is the unobserved individual heterogeneity modeled as individual-specific intercept, which implies that different firms may have different skill levels. u_{it} is the shock with zero mean and variance σ^2. To emphasize the basic idea, the baseline model assumes cross-sectional homoskedasticity, which means that the shock size σ^2 is the same across all firms, which will be relaxed in the general model discussed in Section (5).

As stressed in the motivation, the underlying skill distribution f is the key for better density forecasts. In literature, there are usually two kinds of assumptions imposed on f. One is the random effects (RE) model, where the skill λ_i is independent of the initial performance y_{i0}. The other is the correlated random effects (CRE) model, where the skill λ_i and the initial performance y_{i0} can be potentially correlated with each other. This paper considers both RE and CRE models while focusing on the latter, as the CRE model is more realistic for young firm dynamics as well as many other empirical setups, and RE can be viewed as a special case of CRE with zero correlation.

2.2 Oracle and Feasible Predictors

This subsection formally defines the infeasible optimal oracle predictor and the feasible semiparametric Bayesian predictor proposed in this paper. The kernel of both definitions relies on the conditional

predictor,

$$f_{i,T+1}^{cond}\left(y|\beta,\sigma^2,f,y_{i,0:T}\right)=\int\phi\left(y;\ \beta y_{iT}+\lambda_i,\sigma^2\right)p\left(\lambda_i\left|\beta,\sigma^2,f,y_{i,0:T}\right.\right)d\lambda_i, \tag{2.1}$$

which provides the density forecasts of $y_{i,T+1}$ conditional on the common parameters (β,σ^2), underlying λ_i distribution (f), and firm i's data $(y_{i,0:T})$. The term $\phi\left(y;\ \beta y_{iT}+\lambda_i,\sigma^2\right)$ captures firm i's uncertainty due to future shock, and

$$p\left(\lambda_i\left|\beta,\sigma^2,f,y_{i,0:T}\right.\right)=\frac{p\left(y_{i,1:T}\right|\lambda_i,\beta,\sigma^2,y_{i0})\,f\left(\lambda_i\left|y_{i0}\right.\right)}{\int p\left(y_{i,1:T}\right|\lambda_i,\beta,\sigma^2,y_{i0})\,f\left(\lambda_i\left|y_{i0}\right.\right)d\lambda_i}$$

is the firm-specific posterior that characterizes firm i's uncertainty due to heterogeneous skill. Note that the inference of $\left(\beta,\sigma^2,f\right)$ pools information from the whole cross-section; once conditioned on $\left(\beta,\sigma^2,f\right)$, firms' performances are independent across i, and only firm i's data are needed for its density forecasts.

The infeasible oracle predictor is defined as if we knew all the elements that can be consistently estimated. Specifically, the oracle knows the common parameters (β_0,σ_0^2) and the underlying λ_i distribution (f_0), but not the skill of any individual firm λ_i. Then, the oracle predictor is formulated by plugging the true values $\left(\beta_0,\sigma_0^2,f_0\right)$ into the conditional predictor in equation (2.1),

$$f_{i,T+1}^{oracle}\left(y\right)=f_{i,T+1}^{cond}\left(y|\beta_0,\sigma_0^2,f_0,y_{i,0:T}\right). \tag{2.2}$$

In practice, $\left(\beta,\sigma^2,f\right)$ are all unknown but can be estimated via the Bayesian approach. First, I adopt the conjugate normal-inverse-gamma prior for the common parameters $\left(\beta,\sigma^2\right)$,

$$\left(\beta,\sigma^2\right)\sim N\left(m_0^\beta,\Sigma_0^\beta\right)\mathrm{IG}\left(\sigma^2;\ a_0^{\sigma^2},b_0^{\sigma^2}\right),$$

in order to stay close to the linear Gaussian regression framework. To flexibly model the underlying skill distribution f, I resort to the nonparametric Bayesian prior, which is specified in detail in the next subsection. Then, I update the prior belief using the observations from the whole panel and obtain the posterior. The semiparametric Bayesian predictor is constructed by integrating the conditional predictor over the posterior distribution of $\left(\beta,\sigma^2,f\right)$,

$$f_{i,T+1}^{sp}\left(y\right)=\int f_{i,T+1}^{cond}\left(y|\beta,\sigma^2,f,y_{i,0:T}\right)d\Pi\left(\beta,\sigma^2,f\left|y_{1:N,0:T}\right.\right)d\beta d\sigma^2 df. \tag{2.3}$$

The conditional predictor reflects uncertainties due to future shock and heterogeneous skill, whereas the posterior of $\left(\beta,\sigma^2,f\right)$ captures estimation uncertainty.

2.3 Nonparametric Bayesian Priors

A prior on the skill distribution f can be viewed as a distribution over a set of distributions. Among other options, I choose mixture models for the nonparametric Bayesian prior, because according to the literature, mixture models can effectively approximate a general class of distributions (see Section 4) while being relatively easy to implement (see Section 3). Moreover, the choice of the nonparametric Bayesian prior also depends on whether f is characterized by a random effects model or a correlated random effects model. The correlated random effects setup is more involved but can be crucial in some empirical studies, such as the young firm dynamics application in this paper.

2.3.1 DPM Prior for Random Effects Model

In the random effects model, the skill λ_i is assumed to be independent of the initial performance y_{i0}, so the inference of the underlying skill distribution f can be considered as an unconditional density estimation problem. The DPM model is a typical nonparametric Bayesian prior designed for unconditional density estimation.

Dirichlet Process (DP) The key building block for the DPM model is the DP, which casts a distribution over a set of discrete distributions. A DP has two parameters: the base distribution G_0 characterizing the center of the DP, and the scale parameter α representing the precision (inverse-variance) of the DP. Let G be a distribution drawn from the DP. Denote

$$G \sim DP\left(\alpha, G_0\right),$$

if for any partition $(A_1, \cdots, A_K)$,

$$\left(G\left(A_1\right), \cdots, G\left(A_K\right)\right) \sim \operatorname{Dir}\left(\alpha G_0\left(A_1\right), \cdots, \alpha G_0\left(A_K\right)\right).$$

$\operatorname{Dir}(\cdot)$ stands for the Dirichlet distribution with probability distribution function (pdf) being

$$f_{\operatorname{Dir}}\left(x_1, \cdots, x_K;\ \eta_1, \cdots, \eta_K\right) = \frac{\Gamma\left(\sum_{k=1}^{K} \eta_k\right)}{\prod_{k=1}^{K} \Gamma(\eta_k)} \prod_{k=1}^{K} x_k^{\eta_k - 1},$$

which is a multivariate generalization of the Beta distribution.

An alternative view of DP is given by the stick breaking process,

$$\begin{aligned} G &= \sum_{k=1}^{\infty} p_k \mathbf{1}\left(\theta = \theta_k\right), \\ \theta_k &\sim G_0, \quad k = 1, 2, \cdots, \\ p_k &= \begin{cases} \zeta_1, & k = 1, \\ \prod_{j=1}^{k-1}\left(1-\zeta_j\right)\zeta_k, & k = 2, 3, \cdots, \end{cases} \\ &\text{where } \zeta_k \sim \text{Beta}\left(1,\ \alpha\right), \quad k = 1, 2, \cdots. \end{aligned} \tag{2.4}$$

The stick breaking process distinguishes the roles of G_0 and α in that the former governs component value θ_k while the latter guides the choice of component probability p_k. From now on, for a concise exposition, I denote the p_k part in equation (2.4) as

$$p_k \sim \text{SB}\left(1, \alpha\right),\ k = 1, 2, \cdots,$$

where the function name "SB" is the acronym for "stick breaking", and the two arguments are passed from the parameters of the Beta distribution for "stick length" ζ_k.

Dirichlet Process Mixture (DPM) Prior By definition, a draw from DP is a discrete distribution. In this sense, imposing a DP prior on the skill distribution f amounts to restricting firms' skills to some discrete levels, which may not be very appealing for young firm dynamics as well as some other empirical applications. A natural remedy is to assume λ follows a continuous parametric distribution $f\left(\lambda;\theta\right)$ where θ are the parameters, and adopt a DP prior for the distribution of θ. Then, the parameters θ are discrete while the skill λ enjoys a continuous distribution. This additional layer of mixture lead to the idea of the DPM model. For variables supported on the whole real line, like the skill λ here, a typical choice of the kernel of $f\left(\lambda;\theta\right)$ is a normal distribution with $\theta = \left(\mu, \omega^2\right)$ being the mean and variance of the normal.

$$\begin{aligned} \lambda_i &\sim N\left(\lambda_i;\ \mu_i, \omega_i^2\right), \\ \left(\mu_i, \omega_i^2\right) &\overset{iid}{\sim} G, \\ G &\sim DP\left(\alpha, G_0\right). \end{aligned} \tag{2.5}$$

Equivalently, with component label k, component probability p_k, and component parameters $\left(\mu_k, \omega_k^2\right)$, one draw from the DPM prior can be rewritten as an infinite mixture of normals,

$$\lambda_i \sim \sum_{k=1}^{\infty} p_k N\left(\lambda_i;\ \mu_k, \omega_k^2\right). \tag{2.6}$$

Different draws from the DPM prior are characterized by different combinations of $\{p_k, \mu_k, \omega_k^2\}$, and different combinations of $\{p_k, \mu_k, \omega_k^2\}$ lead to different shapes of f. That is why the DPM prior is flexible enough to approximate many distributions. The component parameters (μ_k, ω_k^2) are directly drawn from the DP base distribution G_0, which is chosen to be the conjugate normal-inverse-gamma distribution. The component probability p_k is constructed via the stick breaking process governed by the DP scale parameter α.

$$\begin{aligned}(\mu_k, \omega_k^2) &\sim G_0, \\ p_k &\sim \mathrm{SB}(1, \alpha), \ k = 1, 2, \cdots.\end{aligned}$$

Comparing the above two sets of expressions in equations (2.5) and (2.6), the first set links the flexible structure in λ to the flexible structure in (μ, ω^2), and serves as a more convenient setup for the theoretical derivation of asymptotic properties as in Subsection 4.3; at the same time, the second set separates the channels regarding component parameters and component probabilities, and therefore is more suitable for the numerical implementation as in Section 3.

One virtue of the nonparametric Bayesian framework is to flexibly elicit the tuning parameter from the data. Namely, we can set up an additional hyperprior for the DP scale parameter α,

$$\alpha \sim \mathrm{Ga}(\alpha;\ a_0^\alpha, b_0^\alpha),$$

and update it based on the observations. Roughly speaking, the DP scale parameter α is linked to the number of unique components in the mixture density and thus determines and reflects the flexibility of the mixture density. Let K^* denote the number of unique components. As derived in Antoniak (1974), we have

$$\begin{aligned}E[K^*|\alpha] &\approx \alpha \log\left(\frac{\alpha + N}{\alpha}\right), \\ Var[K^*|\alpha] &\approx \alpha\left[\log\left(\frac{\alpha + N}{\alpha}\right) - 1\right].\end{aligned}$$

2.3.2 MGLR$_{\mathrm{x}}$ Prior for Correlated Random Effects Model

To accommodate the correlated random effects model where the skill λ_i can be potentially correlated with the initial performance y_{i0}, it is necessary to consider a nonparametric Bayesian prior that is compatible with the much harder conditional density estimation problem. One issue is associated with the uncountable collection of conditional densities, and Pati *et al.* (2013) circumvent it by linking the properties of the conditional density to the corresponding ones of the joint density without explicitly modeling the marginal density of y_{i0}. As suggested in Pati *et al.* (2013), I utilize the Mixtures of Gaussian Linear Regressions (MGLR$_{\mathrm{x}}$) prior, a generalization of the Gaussian-

mixture prior for conditional density estimation. Conditioning on y_{i0},

$$\begin{aligned} \lambda_i | y_{i0} &\sim N\left(\lambda_i;\ \mu_i \left[1, y_{i0}\right]',\ \omega_i^2\right), \\ \left(\mu_i, \omega_i^2\right) &\equiv \theta_i \overset{iid}{\sim} G\left(\cdot;\ y_{i0}\right), \\ G\left(\cdot;\ y_{i0}\right) &= \sum_{k=1}^{\infty} p_k\left(y_{i0}\right) \delta_{\theta_k}. \end{aligned} \tag{2.7}$$

In the baseline setup, both individual heterogeneity λ_i and conditioning set y_{i0} are scalars, so μ_i is a two-element row vector and ω_i^2 is a scalar. Similar to the DPM prior, the component parameters can be directly drawn from the base distribution, which is again specified as the conjugate normal-inverse-gamma distribution,

$$\theta_k \sim G_0, \quad k = 1, 2, \cdots. \tag{2.8}$$

Now the mixture probabilities are characterized by the probit stick breaking process

$$p_k\left(y_{i0}\right) = \Phi\left(\zeta_k\left(y_{i0}\right)\right) \prod_{j<k} \left(1 - \Phi\left(\zeta_j\left(y_{i0}\right)\right)\right), \tag{2.9}$$

where stochastic function ζ_k is drawn from the Gaussian process $\zeta_k \sim GP\left(0, V_k\right)$ for $k = 1, 2, \cdots$.[10]

Expression (2.7) can be perceived as a conditional counterpart of expression (2.5) for the purpose of theoretical derivation. The following expression (2.10) corresponds to expression (2.6), which is in line with the numerical implementation in Section 3:

$$\lambda_i | y_{i0} \sim \sum_{k=1}^{\infty} p_k\left(y_{i0}\right) N\left(\mu_k \left[1, y_{i0}\right]',\ \omega_k^2\right), \tag{2.10}$$

where the component parameters and component probabilities are specified in equations (2.8) and (2.9), respectively.

This setup has three key features: (1) component means are linear in y_{i0}; (2) component variances are independent of y_{i0}; and (3) mixture probabilities are flexible functions of y_{i0}. This framework is general enough to accommodate many conditional distributions. Intuitively, by Bayes' theorem,

$$f\left(\lambda | y_0\right) = \frac{f\left(\lambda, y_0\right)}{f\left(y_0\right)}.$$

[10] For a generic variable c which can be multi-dimensional, the Gaussian process $\zeta\left(c\right) \sim GP\left(m\left(c\right), V\left(c, \tilde{c}\right)\right)$ is defined as follows: for any finite set of $\{c_1, c_2, \cdots, c_n\}$, $\left[\zeta\left(c_1\right), \zeta\left(c_2\right), \cdots, \zeta\left(c_n\right)\right]'$ has a joint Gaussian distribution with the mean vector being $\left[m\left(c_1\right), m\left(c_2\right), \cdots, m\left(c_n\right)\right]'$ and the i,j-th entry of covariance matrix being $V\left(c_i, c_j\right)$, $i, j = 1, \cdots, N$.

The joint distribution in the numerator can be approximated by a mixture of normals

$$f(\lambda, y_0) \approx \sum_{k=1}^{\infty} \tilde{p}_k \phi\left([\lambda, y_0]';\ \tilde{\mu}_k, \tilde{\Omega}_k\right),$$

where $\tilde{\mu}_k$ is a two-element column vector, and $\tilde{\Omega}_k$ is a 2×2 covariance matrix. Applying Bayes' theorem again to the normal kernel for each component k,

$$\phi\left([\lambda, y_0]';\ \tilde{\mu}_k, \tilde{\Omega}_k\right) = \phi\left(y_0;\ \tilde{\mu}_{k,2}, \tilde{\Omega}_{k,22}\right) \phi\left(\lambda;\ \mu_k [1, y_0]', \omega_k^2\right),$$

where $\mu_k = \left[\tilde{\mu}_{k,1} - \frac{\tilde{\Omega}_{k,12}}{\tilde{\Omega}_{k,22}}\tilde{\mu}_{k,2},\ \frac{\tilde{\Omega}_{k,12}}{\tilde{\Omega}_{k,22}}\right]$, $\omega_k^2 = \tilde{\Omega}_{k,11} - \frac{\left(\tilde{\Omega}_{k,12}\right)^2}{\tilde{\Omega}_{k,22}}$. Combining all the steps above, the conditional distribution can be approximated as

$$\begin{aligned} f(\lambda | y_0) &\approx \sum_{k=1}^{\infty} \frac{\tilde{p}_k \phi\left(y_0;\ \tilde{\mu}_{k,2}, \tilde{\Omega}_{k,22}\right) \phi\left(\lambda;\ \mu_k [1, y_0]', \omega_k^2\right)}{f(y_0)} \\ &= \sum_{k=1}^{\infty} p_k(y_0) \phi\left(\lambda;\ \mu_k [1, y_0]', \omega_k^2\right), \end{aligned}$$

The last line is given by collecting marginals of y_{i0} into $p_k(y_0) = \frac{\tilde{p}_k \phi\left(y_0; \tilde{\mu}_{k,2}, \tilde{\Omega}_{k,22}\right)}{f(y_0)}$. In summary, the current setup is similar to approximating the conditional density via Bayes' theorem, but does not explicitly model the distribution of the conditioning variable y_{i0}, and thus allows for more relaxed assumptions on it.

3 Numerical Implementation

In this section, I propose a posterior sampling procedure for the baseline panel data model introduced in Subsection 2.1 together with the nonparametric Bayesian prior specified in Subsection 2.3 that enjoys desirable theoretical properties as discussed in Section 4.

Recall the baseline model,

$$y_{it} = \beta y_{i,t-1} + \lambda_i + u_{it}, \quad u_{it} \sim N\left(0, \sigma^2\right),$$

and the conjugate normal-inverse-gamma prior for the common parameters $\left(\beta, \sigma^2\right)$,

$$\left(\beta, \sigma^2\right) \sim N\left(m_0^{\beta}, \psi_0^{\beta} \sigma^2\right) \mathrm{IG}\left(\sigma^2;\ a_0^{\sigma^2}, b_0^{\sigma^2}\right).$$

The hyperparameters are chosen in a relatively ignorant sense without inferring too much from the data except aligning the scale according to the variance of the data (see Appendix B.1 for details). The skill λ_i is drawn from the underlying skill distribution f, which can be characterized by either

the random effects model or the correlated random effects model. Subsection 3.1 describes the posterior sampler for the former, and Subsection 3.2 delineates the posterior sampler for the latter.

3.1 Random Effects Model

For the random effects model, I impose the Gaussian-mixture DPM prior on f. The posterior sampling algorithm builds on the blocked Gibbs sampler proposed by Ishwaran and James (2001, 2002). They truncate the number of components by a large K, and prove that as long as K is large enough, the truncated prior is "virtually indistinguishable" from the original one. Once truncation is conducted, it is possible to augment the data with latent component probabilities, which boosts numerical convergence and leads to faster code.

To check the robustness regarding the truncation, I also implement the more sophisticated yet complicated slice-retrospective sampler (Dunson, 2009; Yau *et al.*, 2011; Hastie *et al.*, 2015) which does not truncate the number of components at a predetermined K. The full algorithm for the general model (5.1) can be found as Algorithm B.4 in the Appendix. The estimates and forecasts for the two samplers are comparable, so I will only show the results generated from the simpler truncation sampler in this paper.

Suppose the number of components is truncated at K. Then, the Gaussian-mixture DPM prior can be expressed as[11]

$$\lambda_i \sim \sum_{k=1}^{K} p_k N\left(\mu_k, \omega_k^2\right), \quad i = 1, \cdots, N.$$

The parameters for each component can be viewed as directly drawn from the DP base distribution G_0. A typical choice of G_0 is the normal-inverse-gamma prior, which respects the conjugacy when the DPM kernel is also normal (see Appendix B.1 for details of hyperparameter choices).

$$G_0\left(\mu_k, \omega_k^2\right) = N\left(\mu_k;\ m_0^\lambda, \psi_0^\lambda \omega_k^2\right) \mathrm{IG}\left(\omega_k^2;\ a_0^\lambda, b_0^\lambda\right).$$

The component probabilities are constructed via a truncated stick breaking process governed by the DP scale parameter α.

$$p_k = \begin{cases} \zeta_1, & k = 1, \\ \prod_{j=1}^{k-1}\left(1-\zeta_j\right)\zeta_k, & k = 2, \cdots, K-1, \\ 1 - \sum_{j=1}^{K-1} p_j, & k = K, \end{cases}$$
$$\text{where } \zeta_k \sim \mathrm{Beta}\left(1,\ \alpha\right), \quad k = 1, \cdots, K-1.$$

[11] In this section, the nonparametric Bayesian priors are formulated as in equations (2.6) and (2.10). Such expressions explicitly separate the channels regarding component parameters and component probabilities, and hence facilitate the construction of the posterior samplers.

Note that due to the truncation approximation, the probability for component K is different from its infinite mixture counterpart in equation (2.4). Resembling the infinite mixture case, I denote the above truncated sticking process as

$$p_k \sim \text{TSB}\left(1, \alpha, K\right), \ k = 1, \cdots K,$$

where "TSB" is for "truncated stick breaking", the first two arguments are passed from the parameters of the Beta distribution, and the last argument is the truncated number of components.

Let γ_i be firm i's component affiliation, which can take values $\{1, \cdots, K\}$, J_k be the set of firms in component k, i.e. $J_k = \{i : \ \gamma_i = k\}$, and n_k be the number of firms in component k, i.e. $n_k = \#J_k$. Then, the (data-augmented) joint posterior for the model parameters is given by

$$\begin{aligned} & p\left(\alpha, \left\{p_k, \mu_k, \omega_k^2\right\}, \left\{\gamma_i, \lambda_i\right\}, \beta, \sigma^2 \middle| \, y_{1:N,0:T}\right) \\ & = \prod_{i,t} p\left(y_{it} \left| \lambda_i, \beta, \sigma^2, y_{i,t-1}\right.\right) \cdot \prod_i p\left(\lambda_i \left| \mu_{\gamma_i}, \omega_{\gamma_i}^2\right.\right) p\left(\gamma_i \left| \{p_k\}\right.\right) \\ & \quad \cdot \prod_k p\left(\mu_k, \omega_k^2\right) p\left(p_k | \alpha\right) \cdot p\left(\alpha\right) \cdot p\left(\beta, \sigma^2\right), \end{aligned} \tag{3.1}$$

where $k = 1, \cdots, K$, $i = 1, \cdots N$, and $t = 1, \cdots, T$. The first block $\prod_{i,t} p\left(y_{it} \left| \lambda_i, \beta, \sigma^2, y_{i,t-1}\right.\right)$ links observations to model parameters $\left\{\lambda_i\right\}, \beta$, and σ^2. The second block $\prod_i p\left(\lambda_i \left| \mu_{\gamma_i}, \omega_{\gamma_i}^2\right.\right) p\left(\gamma_i \left| \{p_k\}\right.\right)$ links the skill λ_i to the underlying skill distribution f. The last block $\prod_k p\left(\mu_k, \omega_k^2\right) p\left(p_k | \alpha\right) \cdot p\left(\alpha\right) \cdot p\left(\beta, \sigma^2\right)$ formulates the prior belief on $\left(\beta, \sigma^2, f\right)$.

The following Gibbs sampler cycles over the following blocks of parameters (in order): (1) component probabilities, $\alpha, \{p_k\}$; (2) component parameters, $\left\{\mu_k, \omega_k^2\right\}$; (3) component memberships, $\{\gamma_i\}$; (4) individual effects, $\{\lambda_i\}$; (5) common parameters, β, σ^2. A sequence of draws from this algorithm forms a Markov chain with the sampling distribution converging to the posterior density.

Note that if the skill λ_i were known, only step (5) would be sufficient to recover the common parameters. If the mixture structure of f were known (i.e. $\left(p_k, \mu_k, \omega_k^2\right)$ for all components were known), steps (3)-(5) would be needed to first assign firms to components and then infer firm i's skill based on the specific component that it has been assigned to. In reality, neither skill λ_i nor its distribution f is known, so I incorporate two more steps (1)-(2) to model the underlying skill distribution f.

Below, I present the formulas for the key nonparametric Bayesian steps, and leave the details of standard posterior sampling procedures, such as drawing from a normal-inverse-gamma distribution or a linear regression, to Appendix B.3.

Algorithm 3.1. *(Baseline Model: Random Effects)*

For each iteration $s = 1, \cdots, n_{sim}$,

1. *Component probabilities:*

(a) Draw $\alpha^{(s)}$ from a gamma distribution $p\left(\alpha^{(s)}\middle|\, p_K^{(s-1)}\right)$:

$$\alpha^{(s)} \sim Ga\left(\alpha^{(s)};\ a_0^{\alpha} + K - 1,\ b_0^{\alpha} - \log p_K^{(s-1)}\right).$$

(b) For $k = 1, \cdots, K$, draw $p_k^{(s)}$ from the truncated stick breaking process $p\left(\left\{p_k^{(s)}\right\}\middle|\alpha^{(s)}, \left\{n_k^{(s-1)}\right\}\right)$:

$$p_k^{(s)} \sim TSB\left(1 + n_k^{(s-1)},\ \alpha^{(s)} + \sum_{j=k+1}^{K} n_j^{(s-1)},\ K\right),\ k = 1, \cdots K.$$

2. *Component parameters: For $k = 1, \cdots, K$, draw $\left(\mu_k^{(s)}, \omega_k^{2(s)}\right)$ from a normal-inverse-gamma distribution $p\left(\mu_k^{(s)}, \omega_k^{2(s)}\middle|\left\{\lambda_i^{(s-1)}\right\}_{i \in J_k^{(s-1)}}\right)$.*
3. *Component memberships: For $i = 1, \cdots N$, draw $\gamma_i^{(s)}$ from a multinomial distribution $p\left(\left\{\gamma_i^{(s)}\right\}\middle|\left\{p_k^{(s)}, \mu_k^{(s)}, \omega_k^{2(s)}\right\},\ \lambda_i^{(s-1)}\right)$:*

$$\gamma_i^{(s)} = k,\ \textit{with probability}\ p_{ik},\ k = 1, \cdots, K,$$
$$p_{ik} \propto p_k^{(s)} \phi\left(\lambda_i^{(s-1)};\ \mu_k^{(s)}, \omega_k^{2(s)}\right), \quad \sum_{k=1}^{K} p_{ik} = 1.$$

4. *Individual effects: For $i = 1, \cdots, N$, draw $\lambda_i^{(s)}$ from a normal distribution $p\left(\lambda_i^{(s)}\middle|\mu_{\gamma_i^{(s)}}^{(s)}, \omega_{\gamma_i^{(s)}}^{2(s)}, \beta^{(s-1)}, \sigma^{2(s-1)}, y_{i,0:T}\right)$.*
5. *Common parameters: Draw $\left(\beta^{(s)}, \sigma^{2(s)}\right)$ from a linear regression model $p\left(\beta^{(s)}, \sigma^{2(s)}\middle|\left\{\lambda_i^{(s)}\right\},\ y_{1:N,0:T}\right)$.*

3.2 Correlated Random Effects Model

To account for the conditional structure in the correlated random effects model, I implement the MGLR_x prior as specified in Subsection 2.3, which can be viewed as the conditional counterpart of the Gaussian-mixture prior. In the baseline setup, the conditioning set is a singleton with y_{i0} being the only element.

The major computational difference from the random effects model in the previous subsection is that now the component probabilities become flexible functions of y_{i0}. As suggested in Pati *et al.* (2013), I adopt the following priors and auxiliary variables in order to take advantage of conjugacy as much as possible. First, the covariance function for Gaussian process $V_k(c, \tilde{c})$ is specified as

$$V_k(c, \tilde{c}) = \exp\left(-A_k |c - \tilde{c}|^2\right),$$

where $k = 1, 2, \cdots$. An exponential prior is imposed on A_k, i.e.

$$p(A_k) \propto \exp(-A_k),$$

so $p(A_k)$ has full support on $\mathbb{R}^+$ and satisfies Pati *et al.* (2013) Remark 5.2.

Furthermore, it is helpful to introduce a set of auxiliary stochastic functions $\xi_k(y_{i0})$, $k = 1, 2, \cdots$, such that

$$\begin{aligned}
&\xi_k(y_{i0}) \sim N(\zeta_k(y_{i0}), 1), \\
&p_k(y_{i0}) = \text{Prob}(\xi_k(y_{i0}) \geq 0, \text{ and } \xi_j(y_{i0}) < 0 \text{ for all } j < k).
\end{aligned}$$

Note that the probit stick breaking process defined in equation (2.9) can be recovered by marginalizing over $\{\xi_k(y_{i0})\}$.

Finally, I blend the MGLR_x prior with Ishwaran and James (2001, 2002) truncation approximation to simplify the numerical procedure while still retaining reliable results.

Denote $N \times 1$ vectors $\boldsymbol{\zeta}_k = [\zeta_k(y_{10}), \zeta_k(y_{20}), \cdots, \zeta_k(y_{N0})]'$ and $\boldsymbol{\xi}_k = [\xi_k(y_{10}), \xi_k(y_{20}), \cdots, \xi_k(y_{N0})]'$, as well as an $N \times N$ matrix $\boldsymbol{V}_k$ with the ij-th element being $(\boldsymbol{V}_k)_{ij} = \exp\left(-A_k |y_{i0} - y_{j0}|^2\right)$. The next algorithm extends Algorithm 3.1 to the correlated random effects scenario. Step 1 for component probabilities has been changed, while the rest of the steps are in line with those in Algorithm 3.1.

Algorithm 3.2. *(Baseline Model: Correlated Random Effects)*

For each iteration $s = 1, \cdots, n_{sim}$,

1. *Component probabilities:*

   (a) *For* $k = 1, \cdots, K-1$, *draw* $A_k^{(s)}$ *via the random-walk Metropolis-Hastings approach,*

$$p\left(A_k^{(s)} \middle| \boldsymbol{\zeta}_k^{(s-1)}, \{y_{i0}\}\right) \propto \exp\left(-A_k^{(s)}\right) \phi\left(\boldsymbol{\zeta}_k^{(s-1)};\ 0,\ \exp\left(-A_k^{(s)} |y_{i0} - y_{j0}|^2\right)\right).$$

   Then, calculate $\boldsymbol{V}_k^{(s)}$ *such that*

$$\left(\boldsymbol{V}_k^{(s)}\right)_{ij} = \exp\left(-A_k^{(s)} |y_{i0} - y_{j0}|^2\right).$$

   (b) *For* $k = 1, \cdots, K-1$, *and* $i = 1, \cdots, N$, *draw* $\xi_k^{(s)}(y_{i0})$ *from a truncated normal distribution* $p\left(\xi_k^{(s)}(y_{i0}) \middle| \zeta_k^{(s-1)}(y_{i0}), \gamma_i^{(s-1)}\right)$:

$$\xi_k^{(s)}(y_{i0}) \begin{cases} \propto N\left(\zeta_k^{(s-1)}(y_{i0}),\ 1\right) \mathbf{1}\left(\xi_k^{(s)}(y_{i0}) < 0\right), & \text{if } k < \gamma_i^{(s-1)}, \\ \propto N\left(\zeta_k^{(s-1)}(y_{i0}),\ 1\right) \mathbf{1}\left(\xi_k^{(s)}(y_{i0}) \geq 0\right), & \text{if } k = \gamma_i^{(s-1)}, \\ \sim N\left(\zeta_k^{(s-1)}(y_{i0}),\ 1\right), & \text{if } k > \gamma_i^{(s-1)}, \end{cases}$$

(c) *For* $k = 1, \cdots, K-1$, *draw* $\boldsymbol{\zeta}_k^{(s)}$ *from a multivariate normal distribution* $p\left(\boldsymbol{\zeta}_k^{(s)} \middle| \boldsymbol{V}_k^{(s)}, \boldsymbol{\xi}_k^{(s)}\right)$:

$$\begin{aligned}
\boldsymbol{\zeta}_k^{(s)} &\sim N\left(m_k^{\zeta},\ \Sigma_k^{\zeta}\right), \\
\Sigma_k^{\zeta} &= \left[\left(\boldsymbol{V}_k^{(s)}\right)^{-1} + I_N\right]^{-1}, \\
m_k^{\zeta} &= \Sigma_k^{\zeta} \boldsymbol{\xi}_k^{(s)}.
\end{aligned}$$

(d) *For* $k = 1, \cdots, K$, *and* $i = 1, \cdots, N$, *the component probabilities* $p_k^{(s)}(y_{i0})$ *are fully determined by* $\boldsymbol{\zeta}_k^{(s)}$:

$$p_k^{(s)}(y_{i0}) = \begin{cases} \Phi\left(\zeta_1^{(s)}(y_{i0})\right), & \text{if } k = 1, \\ \Phi\left(\zeta_k^{(s)}(y_{i0})\right) \prod_{j<k}\left(1 - \Phi\left(\zeta_j^{(s)}(y_{i0})\right)\right), & \text{if } k = 2, \cdots, K-1, \\ 1 - \sum_{j=1}^{K-1} p_k^{(s)}(y_{i0}), & \text{if } k = K. \end{cases}$$

2. *Component parameters: For* $k = 1, \cdots, K$, *draw* $\left(\mu_k^{(s)}, \omega_k^{2(s)}\right)$ *from a linear regression model* $p\left(\mu_k^{(s)}, \omega_k^{2(s)} \middle| \left\{\lambda_i^{(s-1)}, y_{i0}\right\}_{i \in J_k^{(s-1)}}\right)$.
3. *Component memberships: For* $i = 1, \cdots N$, *draw* $\gamma_i^{(s)}$ *from a multinomial distribution* $p\left(\left\{\gamma_i^{(s)}\right\} \middle| \left\{p_k^{(s)}, \mu_k^{(s)}, \omega_k^{2(s)}\right\},\ \lambda_i^{(s-1)}, y_{i0}\right)$:

$$\begin{aligned}
&\gamma_i^{(s)} = k, \text{ with probability } p_{ik},\ k = 1, \cdots, K, \\
&p_{ik} \propto p_k^{(s)}(y_{i0})\, \phi\left(\lambda_i^{(s-1)};\ \mu_k^{(s)}[1, y_{i0}]', \omega_k^{2(s)}\right), \quad \sum_{k=1}^{K} p_{ik} = 1.
\end{aligned}$$

4. *Individual effects: For* $i = 1, \cdots, N$, *draw* $\lambda_i^{(s)}$ *from a normal distribution* $p\left(\lambda_i^{(s)} \middle| \mu_{\gamma_i^{(s)}}^{(s)}, \omega_{\gamma_i^{(s)}}^{2(s)}, \beta^{(s-1)}, \sigma^{2(s-1)}, y_{i,0:T}\right)$.
5. *Common parameters: Draw* $\left(\beta^{(s)}, \sigma^{2(s)}\right)$ *from a linear regression model* $p\left(\beta^{(s)}, \sigma^{2(s)} \middle| \left\{\lambda_i^{(s)}\right\},\ y_{1:N,0:T}\right)$.

Remark 3.3. With the above prior specification, all steps enjoy closed-form conditional posterior distributions except step 1-a for A_k, which does not exhibit a well-known density form. Hence, I resort to the random-walk Metropolis-Hastings (RWMH) algorithm to sample A_k. In addition, I also incorporate an adaptive procedure based on Atchadé and Rosenthal (2005) and Griffin (2016), which adaptively adjusts the random walk step size and keep acceptance rates around 30%. Intuitively, when the acceptance rate for the current iteration is too high (low), the adaptive algorithm increases (decreases) the step size in the next iteration, and thus potentially raises (lowers) the acceptance rate in the next round. The change in step size decreases with the number of iterations completed, and

the step size converges to the optimal value. Please refer to the detailed description in Algorithm B.1 in the Appendix.

4 Theoretical Properties

4.1 Background

Generally speaking, Bayesian analysis starts with a prior belief and updates it with data. It is desirable to ensure that the prior belief does not dominate the posterior inference asymptotically. Namely, as more and more data have been observed, one would have weighed more on the data and less on prior, and the effect from the prior would have ultimately been washed out. For pure Bayesians who have different prior beliefs, the asymptotic properties make sure that they will eventually agree on similar predictive distributions (Blackwell and Dubins, 1962; Diaconis and Freedman, 1986). For frequentists who perceive that there is an unknown true data generating process, the asymptotic properties act as frequentist justification for the Bayesian analysis—as the sample size increases, the updated posterior recovers the unknown truth. Moreover, the conditions for posterior consistency provide guidance in choosing better-behaved priors.

In the context of infinite dimensional analysis such as density estimation, posterior consistency cannot be taken as given. On the one hand, Doob's theorem (Doob, 1949) indicates that Bayesian posterior will achieve consistency almost surely under the prior measure. On the other hand, the null set for the prior can be topologically large, and hence the true model can easily fall beyond the scope of the prior, especially in nonparametric analysis. Freedman (1963) gives a simple counter-example in the nonparametric setup, and Freedman (1965) further examines the combinations of the prior and the true parameters that yield a consistent posterior, and proves that such combinations are meager in the joint space of the prior and the true parameters. Therefore, for problems involving density estimation, it is crucial to find reasonable conditions on the joint behavior of the prior and the true density to establish the posterior consistency argument.

In this section, I show the asymptotic properties of the proposed semiparametric Bayesian predictor when the time dimension T is fixed and the cross-sectional dimension N tends to infinity. Basically, under reasonably general conditions, the joint posterior of the common parameters and the individual effect distribution concentrates in an arbitrarily small region around the true data generating process, and the density forecasts concentrate in an arbitrarily small region around the oracle. Subsection 4.2 provides the conditions for identification, which lays the foundation for posterior consistent analysis. Subsection 4.3 proves the posterior consistency of the estimator, which also serves as an essential building block for bounding the discrepancy between the proposed predictor and the oracle. Finally, Subsection 4.4 establishes the Bayesian asymptotic argument for density forecasts.

4.2 Identification

To establish the posterior consistency argument, we first need to ensure identification for both the common parameters and the (conditional) distribution of individual effects. Here, I present the identification result in terms of the correlated random effects model, with the random effects model being a special case. In the baseline setup, the identification argument directly follows Assumptions 2.1-2.2 and Theorem 2.3 in Liu *et al.* (2016), which is in turn based on early works, such as Arellano and Bover (1995) and Arellano and Bonhomme (2012), so below I only state the assumption and the proposition without extensive discussion. For more general results addressing correlated random coefficients, cross-sectional heteroskedasticities, and unbalanced panels, please refer to Subsection 5.3.

Assumption 4.1. *(Baseline Model: Identification)*

1. $\{y_{i0},\ \lambda_i\}$ *are i.i.d. across* i.
2. u_{it} *is i.i.d. across* i *and* t, *and independent of* λ_i.
3. *The characteristic function for* $\lambda_i|y_{i0}$ *is non-vanishing almost everywhere.*
4. $T \geq 2$.

The first condition characterizes the correlated random effects model, where there can be potential correlation between skill λ_i and initial performance y_{i0}. For the random effects case, this condition can be altered to "λ_i is independent of y_{i0} and i.i.d. across i". The second condition implies that skill is independent of shock, and that shock is independent across firms and times, so skill and shock are intrinsically different and distinguishable. The third condition facilitates the deconvolution between the signal (skill) and the noise (shock) via Fourier transformation. The last condition guarantees that the time span is long enough to distinguish persistence ($\beta y_{i,t-1}$) and individual effects (λ_i). Then, the identification statement is established as follows.

Proposition 4.2. *(Baseline Model: Identification)*

Under Assumption 4.1, the common parameters $\left(\beta, \sigma^2\right)$ *and the conditional distribution of individual effects* $f(\lambda_i|y_{i0})$ *are all identified.*

4.3 Posterior Consistency

In this subsection, I establish the posterior consistency of the estimated common parameters $\left(\beta, \sigma^2\right)$ and the estimated (conditional) distribution of individual effects f in the baseline setup. Note that the estimated individual effects λ_is are not consistent because information is accumulated only along the cross-section dimension but not along the time dimension. Subsections 4.3.1 and 4.3.2 examine the random effects model and the correlated random effects model, respectively. Further discussion of the general model can be found in Subsection 5.4.

4.3.1 Random Effects Model

First, let us consider the random effects model with f being an unconditional distribution. Let $\Theta = \mathbb{R} \times \mathbb{R}^+$ be the space of the parametric component $\vartheta = \left(\beta, \sigma^2\right)$, and let $\mathcal{F}$ be the set of densities on $\mathbb{R}$ (with respect to Lebesgue measure) as the space of the nonparametric component f. The true data generating process is characterized by (ϑ_0, f_0). The posterior consistency results are established with respect to the weak topology, which is generated by a neighborhood basis constituted of the weak neighborhoods defined below and is closely related to convergence in distribution or weak convergence.

Definition 4.3. A *weak neighborhood* of f_0 is defined as

$$U_{\epsilon,\Phi}(f_0) = \left\{ f \in \mathcal{F} : \left| \int \varphi_j f - \int \varphi_j f_0 \right| < \epsilon \right\}$$

where $\epsilon > 0$ and $\Phi = \{\varphi_j\}_{j=1}^J$ are bounded, continuous functions.

Let $\Pi(\cdot, \cdot)$ be a joint prior distribution on $\Theta \times \mathcal{F}$ with marginal priors being $\Pi^{\vartheta}(\cdot)$ and $\Pi^f(\cdot)$. The corresponding joint posterior distribution is denoted as $\Pi(\cdot, \cdot | y_{1:N,0:T})$ with the marginal posteriors indicated with superscripts.

Definition 4.4. The posterior achieves *weak consistency* at (ϑ_0, f_0) if for any $U_{\epsilon,\Phi}(f_0)$ and any $\delta > 0$, as $N \to \infty$,

$$\Pi\left((\vartheta, f) : \ \|\vartheta - \vartheta_0\| < \delta, \ f \in U_{\epsilon,\Phi}(f_0) \middle| \, y_{1:N,0:T} \right) \to 1, \ a.s.$$

As stated in the original Schwartz (1965) theorem (Lemma 4.6), weak consistency is closely related to the Kullback-Leibler (KL) divergence. For any two distributions f_0 and f, the *KL divergence* of f from f_0 is defined as

$$d_{KL}(f_0, f) = \int f_0 \log \frac{f_0}{f}.$$

The *KL property* is characterized based on KL divergence as follows.

Definition 4.5. If for all $\epsilon > 0$, $\Pi^f(f \in \mathcal{F} : \ d_{KL}(f_0, f) < \epsilon) > 0$, we say f_0 is *in the KL support of* Π^f, or $f_0 \in KL\left(\Pi^f\right)$.

Preliminary: Schwartz (1965) Theorem The following lemma restates the Schwartz (1965) theorem of weak posterior consistency. It is established in a simpler scenario where we observe λ_i (not y_i) and wants to infer its distribution.

Lemma 4.6. *(Schwartz, 1965)*

The posterior is weakly consistent at f_0 under two sufficient conditions:

1. *Kullback-Leibler property: f_0 is in the KL support of Π, or $f_0 \in KL(\Pi)$.*
2. *Uniformly exponentially consistent tests: For any $U = U_{\epsilon,\Phi}(f_0)$, there exists $\gamma > 0$ and a sequence of tests $\boldsymbol{\varphi_N}(\lambda_1, \cdots, \lambda_N)$ testing*[12]

$$H_0 : \ f = f_0 \quad against \quad H_1 : \ f \in U^c$$

such that[13]

$$\mathbb{E}_{f_0}(\boldsymbol{\varphi_N}) < \exp(-\gamma N) \quad and \quad \sup_{f \in U^c} \mathbb{E}_f(1 - \boldsymbol{\varphi_N}) < \exp(-\gamma N) \tag{4.1}$$

for all $N > N_0$, where N_0 is a positive integer.

The following sketch of proof gives the intuition behind the two sufficient conditions. Note that the posterior probability of U^c is given by

$$\begin{aligned} \Pi(U^c|\lambda_{1:N}) &= \frac{\int_{U^c} \prod_{i=1}^N \frac{f(\lambda_i)}{f_0(\lambda_i)} d\Pi(f)}{\int_{\mathcal{F}} \prod_{i=1}^N \frac{f(\lambda_i)}{f_0(\lambda_i)} d\Pi(f)} \equiv \frac{\text{numer}_N}{\text{denom}_N} \\ &\leq \boldsymbol{\varphi_N} + \frac{(1 - \boldsymbol{\varphi_N}) \text{numer}_N}{\text{denom}_N}, \end{aligned} \tag{4.2}$$

and we want it to be arbitrarily small.

First, based on the Borel-Cantelli lemma, the condition on the type-I error suggests that the first term $\boldsymbol{\varphi_N} \to 0$ almost surely.

Second, for the numerator of the second term, the condition on the type-II error implies that

$$\begin{aligned} \mathbb{E}_{f_0}((1 - \boldsymbol{\varphi_N}) \text{numer}_N) &= \int (1 - \boldsymbol{\varphi_N}) \cdot \int_{U^c} \prod_{i=1}^N \frac{f(\lambda_i)}{f_0(\lambda_i)} d\Pi(f) \cdot \prod_{i=1}^N f_0(\lambda_i) \, d\lambda_i \\ &= \int_{U^c} \int (1 - \boldsymbol{\varphi_N}) \prod_{i=1}^N f(\lambda_i) \, d\lambda_i \cdot d\Pi(f) \\ &\leq \sup_{f \in U^c} \mathbb{E}_f((1 - \boldsymbol{\varphi_N})) \\ &< \exp(-\gamma N). \end{aligned}$$

Hence, $\exp\left(\frac{\gamma N}{2}\right)(1 - \boldsymbol{\varphi_N}) \text{numer}_N \to 0$ almost surely.

Third, for the denominator of the second term, as $N \to 0$,

$$\text{denom}_N = \int_{\mathcal{F}} \exp\left(-\sum_{i=1}^N \log \frac{f_0(\lambda_i)}{f(\lambda_i)}\right) d\Pi(f) \to \int_{\mathcal{F}} \exp(-N \cdot d_{KL}(f_0, f)) \, d\Pi(f).$$

[12] $\boldsymbol{\varphi_N} = 0$ favors the null hypothesis H_0, whereas $\boldsymbol{\varphi_N} = 1$ favors the alternative hypothesis H_1.

[13] $\mathbb{E}_{f_0}(\boldsymbol{\varphi_N})$ and $\sup_{f \in U^c} \mathbb{E}_f(1 - \boldsymbol{\varphi_N})$ can be interpreted as type-I and type-II errors, respectively.

Combine it with the KL property $f_0 \in KL(\Pi)$, then

$$\liminf_{N\to\infty} e^{\tilde{\gamma}N} \cdot \text{denom}_N = \infty, \text{ for all } \tilde{\gamma} > 0.$$

Hence, $\exp\left(\frac{\gamma N}{4}\right) \text{denom}_N \to \infty$ almost surely.

Therefore, the posterior probability of U^c

$$\Pi(U^c|\lambda_{1:N}) \to 0, \ a.s.$$

Schwartz (1965) Theorem guarantees posterior consistency in a general density estimation context. However, as mentioned in the introduction, there are a number of challenges in adapting these two conditions even to the baseline setup with random effects. The first challenge is that, because we observe y_{it} rather than λ_i, we need to disentangle the uncertainties generated from unknown cross-sectional heterogeneities λ_is and from independent shocks u_{it}s. Second is to incorporate unknown shock size σ^2. Third is to take care of the lagged dependent variables as covariates.

In all these scenarios, note that:

(1) The KL requirement ensures that the prior puts positive weight on the true distribution. To satisfy the KL requirement, we need some joint assumptions on the true distribution f_0 and the prior Π. Compared to general nonparametric Bayesian modeling, the DPM structure (and the MGLR_x structure for the correlated random effects model) offers more regularities on the prior Π and thus weaker assumptions on the true distribution f_0 (see Lemma 4.8 and Assumption 4.14).

(2) Uniformly exponentially consistent tests guarantee that the data is informative enough to differentiate the true distribution from the alternatives. These tests are not specific to the DPM setup but closely related to the definition of the weak neighborhood, hence linked to the identification argument as well.

In the following discussion, I will tackle the aforementioned three challenges one by one.

Disentangle Skills and Shocks Now let us consider a simple cross-sectional case where $\beta = 0$, $\sigma^2 = 1$, and $T = 1$. Since there is only one period, the t subscript is omitted.

$$y_i = \lambda_i + u_i, \quad u_i \sim N(0,1), \tag{4.3}$$

The only twist here is to distinguish the uncertainties originating from unknown individual effects λ_is and from independent shocks u_is. Note that unlike previous studies that estimate distributions of observables,[14] here the target λ_i intertwines with u_i and cannot be easily inferred from the observed y_i, i.e. a deconvolution problem.

[14]Some studies (Amewou-Atisso *et al.*, 2003; Tokdar, 2006) estimate distributions of quantities that can be inferred from observables given common coefficients. For example, in the linear regression problems with an unknown error distribution, i.e. $y_i = \beta' x_i + u_i$, conditional on the regression coefficients β, $u_i = y_i - \beta' x_i$ is inferable from the data.

Proposition 4.7. *(Baseline Model: Skills vs Shocks)*

In setup (4.3) with the random effects version of Assumption 4.1 (1-3), if $f_0 \in KL\left(\Pi^f\right)$*, the posterior is weakly consistent at* f_0*.*

At the first glance, Proposition 4.7 looks similar to the classical Schwartz (1965) theorem. However, here both the KL requirement and the uniformly exponentially consistent tests are constructed on the observed y_i whereas the weak consistency result is established on the unobserved λ_i. There is a gap between the two, as previously mentioned.

The KL requirement is achieved through the convexity of the KL divergence. In terms of the tests, intuitively, if we obtain enough data and know the distribution of the shocks, it is possible to separate the signal λ_i from the noise u_i even in the cross-sectional setting. The exact argument is delivered via proof by contradiction that utilizes characteristic functions to uncouple the effects from λ_i and u_i. Please refer to Appendix C.1.1 for the detailed proof.

Previous studies have proposed many sets of sufficient conditions to ensure that f_0 is in the KL support of Π^f. Based on Wu and Ghosal (2008) Theorem 5, the next lemma gives one set of sufficient conditions for f_0 together with the Gaussian-mixture DPM prior,[15]

$$\begin{aligned} \lambda_i &\sim N\left(\mu_i, \omega_i^2\right), \\ \left(\mu_i, \omega_i^2\right) &\overset{iid}{\sim} G, \\ G &\sim DP\left(\alpha, G_0\right). \end{aligned}$$

Lemma 4.8. *(Wu and Ghosal, 2008: Gaussian)*

If f_0 *and its prior* G_0 *satisfy the following conditions:*

1. $f_0(\lambda)$ *is a continuous density on* $\mathbb{R}$*.*
2. *For some* $0 < M < \infty$*,* $0 < f_0(\lambda) \leq M$ *for all* λ*.*
3. $\left|\int f_0(\lambda) \log f_0(\lambda)\, d\lambda\right| < \infty$*.*
4. *For some* $\delta > 0$*,* $\int f_0(\lambda) \log \frac{f_0(\lambda)}{\varphi_\delta(\lambda)} d\lambda < \infty$*, where* $\varphi_\delta(\lambda) = \inf_{\|\lambda' - \lambda\| < \delta} f_0(\lambda')$*.*
5. *For some* $\eta > 0$*,* $\int |\lambda|^{2(1+\eta)} f_0(\lambda)\, d\lambda < \infty$*.*
6. G_0 *has full support on* $\mathbb{R} \times \mathbb{R}^+$*.*

then $f_0 \in KL\left(\Pi^f\right)$*.*

Conditions 1-5 ensure that the true distribution f_0 is well-behaved, and condition 6 further guarantees that the DPM prior is general enough to contain the true distribution.

If the true distribution f_0 has heavy tails, we can resort to Lemma E.1 following Tokdar (2006) Theorem 3.3. Lemma E.1 ensures the posterior consistency of Cauchy f_0 when G_0 is the standard conjugate normal-inverse-gamma distribution.

[15] In this section, the nonparametric Bayesian priors are in the form of equations (2.5) and (2.7), which are more suitable for the posterior consistency analysis.

Unknown Shock Size Most of the time in practice, we do not know the shock variances in advance. In this section, I consider cross-sectionally homoskedastic shocks with unknown variance as in the baseline model. The cross-sectional heteroskedasticity scenario can be found in Subsection 5.4.1. Now consider a panel setting $(T > 1)$[16] with $\beta = 0$:

$$y_{it} = \lambda_i + u_{it}, \quad u_{it} \sim N\left(0, \sigma^2\right), \tag{4.4}$$

where σ^2 is unknown with the true value being σ_0^2. The joint posterior consistency for $\left(\sigma^2, f\right)$ is stated in the following proposition.

Proposition 4.9. *(Baseline Model: Unknown Shock Size)*

In setup (4.4) with the random effects version of Assumption 4.1, if $f_0 \in KL\left(\Pi^f\right)$ *and* $\sigma_0^2 \in supp\left(\Pi^{\sigma^2}\right)$*, the posterior is weakly consistent at* $\left(\sigma_0^2, f_0\right)$*.*

Paralleling the previous subsection, we can refer to Lemma 4.8 for sufficient conditions that ensure $f_0 \in KL\left(\Pi^f\right)$.

Appendix C.1.2 provides the complete proof. The KL requirement is satisfied based on the dominated convergence theorem. The intuition behind the tests is to split the alternative region of $\left(\sigma^2, f\right)$ into two parts. First, when a candidate σ^2 is far from the true σ_0^2, we can employ orthogonal forward differencing to get rid of λ_i (see Appendix D.1), and then use the residues to construct a sequence of tests which distinguish Gaussian distributions with different variances. Second, when σ^2 is close to σ_0^2 but f is far from f_0, we need to make sure that the deviation generated from σ^2 is small enough so that it cannot offset the difference in f.

Lagged Dependent Variables Lagged dependent variables are essential for most economic predictions, as persistence is usually an important feature of economic data. Now let us add a one-period lag of y_{it} to the right hand side of equation (5.4), which gives exactly the baseline model (1.1):

$$y_{it} = \beta y_{i,t-1} + \lambda_i + u_{it}, \quad u_{it} \sim N\left(0, \sigma^2\right),$$

where $\vartheta = \left(\beta, \sigma^2\right)$ are unknown with the true value being $\vartheta_0 = \left(\beta_0, \sigma_0^2\right)$. The following assumption ensures the existence of the required tests in the presence of a linear regressor.

Assumption 4.10. *(Initial Conditions)*

y_{i0} *is compactly supported.*

Proposition 4.11. *(Baseline Model: Random Effects)*

In the baseline setup (1.1) with random effects, suppose we have:

[16] Note that when λ_i and u_{it} are both Gaussian with unknown variances, we cannot separately identify the variances in the cross-sectional setting $(T = 1)$. This is no longer a problem if either of the distributions is non-Gaussian or if we work with panel data.

1. *The random effects version of Assumption 4.1.*
2. *y_{i0} satisfies Assumption 4.10.*
3. *f and G satisfy Lemma 4.8.*
4. $\vartheta_0 \in \operatorname{supp}\left(\Pi^{\vartheta}\right)$.

Then, the posterior is weakly consistent at (ϑ_0, f_0).

The proof can be found in Appendix C.1.3. The KL requirement is established as in previous cases. The uniformly exponentially consistent tests are constructed by dividing the alternative region into two parts: the tests on β and σ^2 are achieved via orthogonal forward differencing followed by a linear regression, while the tests on f are crafted to address the non-i.i.d. observables due to the AR(1) term.

Once again, we can refer to Tokdar (2006) Theorem 3.3 in order to account for heavy tails in the true unknown distributions. For further details, please see Proposition E.3 regarding the general model (5.1).

4.3.2 Correlated Random Effects Model

In the young firm example, the correlated random effects model can be interpreted as that a young firm's initial performance may reflect its underlying skill, which is a more sensible assumption.

For the correlated random effects model, the definitions and notations are parallel with the random effects ones with slight adjustment considering that now f is a conditional distribution. In the baseline setup, the conditioning set $c_i = y_{i0}$. As in Pati *et al.* (2013), it is helpful to link the properties of the conditional density to the corresponding ones of the joint density without explicitly modeling the marginal density of y_{i0}, which circumvents the difficulty associated with an uncountable set of conditional densities. Let $\mathcal{C}$ be a compact subset of $\mathbb{R}$ for the conditioning variable $c_i = y_{i0}$, $\mathcal{H}$ be the set of joint densities on $\mathbb{R} \times \mathcal{C}$ (with respect to Lebesgue measure), and $\mathcal{F}$ be the set of conditional densities on $\mathbb{R}$ given conditioning variable $c \in \mathcal{C}$.

Let h, f, and q be the joint, conditional, and marginal densities, respectively. Denote

$$h_0(\lambda, c) = f_0(\lambda|c) \cdot q_0(c), \quad h(\lambda, c) = f(\lambda|c) \cdot q_0(c).$$

where $h,\ h_0 \in \mathcal{H}$, and $f,\ f_0 \in \mathcal{F}$. h_0, f_0, and q_0 are the true densities. Note that h and h_0 share the same marginal density q_0, but different conditional densities f and f_0. This setup does not require estimating q_0 and thus relaxes the assumption on the initial conditions.

The definitions of weak neighborhood and KL property rely on this joint density characterization. Note that in both definitions, the conditioning variable c is integrated out with respect to the true q_0.

Definition 4.12. A *weak neighborhood* of f_0 is defined as

$$U_{\epsilon,\Phi}(f_0) = \left\{ f \in \mathcal{F} : \left| \int \varphi_j h - \int \varphi_j h_0 \right| < \epsilon \right\}$$

where $\epsilon > 0$ and $\Phi = \{\varphi_j\}_{j=1}^J$ are bounded, continuous functions of (λ, c).

Definition 4.13. If for all $\epsilon > 0$, $\Pi^f (f \in \mathcal{F} : d_{KL}(h_0, h) < \epsilon) > 0$, we say f_0 is *in the KL support of* Π^f, or $f_0 \in KL\left(\Pi^f\right)$.

As described in Subsection 2.3.2, the MGLR_x prior is a conditional version of the nonparametric Bayesian prior. It can be specified as follows, with the conditioning set simply being a scalar, y_{i0}.

$$\begin{aligned} \lambda_i | y_{i0} &\sim N\left(\lambda_i;\ \mu_i [1, y_{i0}]',\ \omega_i^2\right), \\ \left(\mu_i, \omega_i^2\right) \equiv \theta_i &\overset{iid}{\sim} G(\cdot;\ y_{i0}), \\ G(\cdot;\ y_{i0}) &= \sum_{k=1}^{\infty} p_k(y_{i0})\, \delta_{\theta_k}. \end{aligned}$$

where for components $k = 1, 2, \cdots$

$$\begin{aligned} \theta_k &\sim G_0, \\ p_k(y_{i0}) &= \Phi(\zeta_k(y_{i0})) \prod_{j<k} (1 - \Phi(\zeta_j(y_{i0}))), \\ \zeta_k &\sim GP(0, V_k). \end{aligned}$$

The induced prior on the mixing measures $G(\theta_i;\ y_{i0})$ is denoted as $\tilde{\Pi}$.

Assumption 4.14. *(Baseline Model: Correlated Random Effects)*

1. *Conditions on f_0:*
   (a) *For some $0 < M < \infty$, $0 < f_0(\lambda|y_0) \leq M$ for all (λ, y_0).*
   (b) $\left| \int \left[\int f_0(\lambda|y_0) \log f_0(\lambda|y_0)\, d\lambda \right] q_0(y_0)\, dy_0 \right| < \infty$.
   (c) $\left| \int \left[\int f_0(\lambda|y_0) \log \frac{f_0(\lambda|y_0)}{\varphi_\delta(\lambda|y_0)} d\lambda \right] q_0(y_0)\, dy_0 \right| < \infty$, *where* $\varphi_\delta(\lambda|y_0) = \inf_{|\lambda' - \lambda| < \delta} f_0(\lambda|y_0)$, *for some $\delta > 0$.*
   (d) *For some $\eta > 0$,* $\int \left[\int |\lambda|^{2(1+\eta)} f_0(\lambda|y_0)\, d\lambda \right] q_0(y_0)\, dy_0 < \infty$.
   (e) *$f_0(\cdot|\cdot)$ is jointly continuous in (λ, y_0).*
   (f) *$q_0(y_0) > 0$ for all $y_0 \in \mathcal{C}$.*
2. *Conditions on $\tilde{\Pi}$:*
   (a) *For $k = 1, 2, \cdots$, V_k is chosen such that $\zeta_k \sim GP(0, V_k)$ has continuous path realizations.*
   (b) *For $k = 1, 2, \cdots$, for any continuous $g(\cdot)$, and any $\epsilon > 0$,* $\tilde{\Pi}\left(\sup_{y_0 \in \mathcal{C}} |\zeta_k(y_0) - g(y_0)| < \epsilon\right) > 0$.
   (c) *G_0 is absolutely continuous.*

These conditions follow Assumptions A1-A5 and S1-S3 in Pati *et al.* (2013) for posterior consistency under the conditional density topology. The first group of conditions can be viewed as conditional density analogs of the conditions in Lemma 4.8. These requirements are satisfied for flexible classes of models, i.e. generalized stick-breaking process mixtures with the stick-breaking lengths being monotone differentiable functions of a continuous stochastic process.

Proposition 4.15. *(Baseline Model: Correlated Random Effects)*

In the baseline setup (1.1) with correlated random effects, suppose we have:

1. *Assumption 4.1.*
2. y_{i0} *satisfies Assumption 4.10.*
3. f *and* G *satisfy Assumption 4.14.*
4. $\vartheta_0 \in \operatorname{supp}\left(\Pi^{\vartheta}\right)$.

Then, the posterior is weakly consistent at (ϑ_0, f_0).

The proof in Appendix C.2 is similar to the random effects case except that now both the KL property and the uniformly exponentially consistent tests are constructed on h versus h_0 rather than f versus f_0.

4.4 Density forecasts

Once the posterior consistency results are obtained, we can bound the discrepancy between the proposed predictor and the oracle by the estimation uncertainties in β, σ^2, and f, and then show the asymptotical convergence of the density forecasts to the oracle forecast (see Appendix C.3 for the detailed proof).

Proposition 4.16. *(Baseline Model: Density Forecasts)*

In the baseline setup (1.1), suppose we have:

1. *For the random effects model, conditions in Proposition 4.11.*
2. *For the correlated random effects model,*
   (a) *conditions in Proposition 4.15,*
   (b) $q_0(y_0)$ *is continuous, and there exists* $\underline{q} > 0$ *such that* $|q_0(y_0)| > \underline{q}$ *for all* $y_0 \in \mathcal{C}$.

Then, the density forecasts converge to the oracle predictor in the following two ways:

1. *Convergence of* $f^{cond}_{i,T+1}$ *in weak topology: for any* i *and any* $U_{\epsilon,\Phi}\left(f^{oracle}_{i,T+1}\right)$, *as* $N \to \infty$,

$$\mathbb{P}\left(f^{cond}_{i,T+1} \in U_{\epsilon,\Phi}\left(f^{oracle}_{i,T+1}\right) \middle| y_{1:N,0:T}\right) \to 1, \ a.s.$$

2. *"Pointwise" convergence of* $f^{sp}_{i,T+1}$*: for any* i, *any* y, *and any* $\epsilon > 0$, *as* $N \to \infty$,

$$\left|f^{sp}_{i,T+1}(y) - f^{oracle}_{i,T+1}(y)\right| < \epsilon, \ a.s.$$

The first result focuses on the conditional predictor (2.1) and is more coherent with the weak topology for posterior consistency in the previous subsection. The second result is established for the semiparametric Bayesian predictor (2.3), which is the posterior mean of the conditional predictor. In addition, the asymptotic convergence of aggregate-level density forecasts can be derived by summing individual-specific forecasts over different subcategories.

5 Extensions

5.1 General Panel Data Model

The general panel data model with correlated random coefficients can be specified as

$$y_{it} = \beta' x_{i,t-1} + \lambda_i' w_{i,t-1} + u_{it}, \quad u_{it} \sim N\left(0, \sigma_i^2\right) \tag{5.1}$$

where $i = 1, \cdots, N$, and $t = 1, \cdots, T+1$. Similar to the baseline setup in Subsection 2.1, the y_{it} is the observed individual outcomes, and I am interested in providing density forecasts of $y_{i,T+1}$ for any individual i.

The $w_{i,t-1}$ is a vector of observed covariates that have heterogeneous effects on the outcomes, with λ_i being the unobserved individual heterogeneities. $w_{i,t-1}$ is strictly exogenous and captures the key sources of individual heterogeneities. The simplest choice would be $w_{i,t-1} = 1$ where λ_i can be interpreted as an individual-specific intercept, i.e. firm i's skill level in the baseline model (1.1). Moreover, it is also helpful to include other key covariates of interest whose effects are more diverse cross-sectionally, such as observables that characterize innovation activities. Furthermore, the current setup can also take into account deterministic or stochastic aggregate effects, such as time dummies for the recent recession. For notation clarity, I decompose $w_{i,t-1} = \left(w_{t-1}^{A\prime}, w_{i,t-1}^{I\prime}\right)'$, where w_{t-1}^{A} stands for a vector of aggregate variables, and $w_{i,t-1}^{I}$ is composed of individual-specific variables.

The $x_{i,t-1}$ is a vector of observed covariates that have homogeneous effects on the outcomes, and β is the corresponding vector of common parameters. $x_{i,t-1}$ can be either strictly exogenous or predetermined, which can be further denoted as $x_{i,t-1} = \left(x_{i,t-1}^{O\prime}, x_{i,t-1}^{P\prime}\right)'$, where $x_{i,t-1}^{O}$ is the strictly exogenous part while $x_{i,t-1}^{P}$ is the predetermined part. The one-period-lagged outcome $y_{i,t-1}$ is a typical candidate for $x_{i,t-1}^{P}$ in the dynamic panel data literature, which captures the persistence structure. In addition, both $x_{i,t-1}^{O}$ and $x_{i,t-1}^{P}$ can incorporate other general control variables, such as firm characters as well as local and national economic conditions. The notation $x_{i,t-1}^{P*}$ indicates the subgroup of $x_{i,t-1}^{P}$ excluding lagged outcomes. Here, the distinction between homogeneous effects ($\beta' x_{i,t-1}$) versus heterogeneous effects ($\lambda_i' w_{i,t-1}$) allows us to enjoy the best of both worlds—revealing the latent nonstandard structures for the key effects while avoiding the curse-of-dimensionality problem, which shares the same idea as Burda *et al.* (2012).

The u_{it} is an individual-time-specific shock characterized by zero mean and cross-sectional heteroskedasticity, σ_i^2. The normality assumption is not very restrictive due to the flexibility in σ_i^2 distribution. Table 1 in Fernandez and Steel (2000) demonstrates that scale mixture of normals can capture "a rich class of continuous, symmetric, and unimodal distributions" (p. 81), including Cauchy, Laplace, Logistic, etc. More rigorously, as proved by Kelker (1970), this class is composed of marginal distributions of higher-dimensional spherical distributions.

In the correlated random coefficients model, λ_i can depend on some of the covariates and initial conditions. Specifically, I define the conditioning set at period t to be

$$c_{i,t-1} = \left\{ y_{i,0:t-1},\ x^{P*}_{i,0:t-1},\ x^{O}_{i,0:T},\ w_{i,0:T} \right\} \tag{5.2}$$

and allow the distribution of λ_i and σ_i^2 to be a function of c_{i0}. Note that as lagged y_{it} and $x^{P*}_{i,t-1}$ are predetermined variables, the sequences of $x^{P*}_{i,t-1}$ in the conditioning set $c_{i,t-1}$ start from period 0 to period $t-1$; while $x^{O}_{i,t-1}$ and $w_{i,t-1}$ are both strictly exogenous, so the conditioning set $c_{i,t-1}$ contains their entire sequences. For future use, I also define the part of $c_{i,t-1}$ that is composed of individual-specific variables as

$$c^*_{i,t-1} = \left\{ y_{i,0:t-1},\ x^{P*}_{i,0:t-1},\ x^{O}_{i,0:T},\ w^{I}_{i,0:T} \right\}.$$

Furthermore, the above setup can be extended to unbalanced panels. Let T_i denote the longest chain for individual i that has complete observations, from t_{0i} to t_{1i}. That is, $\{y_{it},\ w_{i,t-1},\ x_{i,t-1}\}$ are observed for all $t = t_{0i}, \cdots, t_{1i}$. Then, I discard the unobserved periods and redefine the conditioning set at time $t = 1, t_{0i}, \cdots, t_{1i}, T+1$ to be

$$c_{i,t-1} = \left\{ y_{i,\tau^P_{i,t-1}},\ x^{P*}_{i,\tau^P_{i,t-1}},\ x^{O}_{i,\tau^P_{iT}},\ w_{i,\tau^P_{iT}} \right\}, \tag{5.3}$$

where the set for time periods $\tau^P_{i,t-1} = \{0, t_{0i}-1, \cdots, t_{1i}-1, T\} \cap \{0, \cdots, t-1\}$. Note that t_{i0} can be 1, and t_{i1} can be T, so this structure is also able to accommodate balanced panels. Accordingly, the individual-specific component of $c_{i,t-1}$ is

$$c^*_{i,t-1} = \left\{ y_{i,\tau^P_{i,t-1}},\ x^{P*}_{i,\tau^P_{i,t-1}},\ x^{O}_{i,\tau^P_{iT}},\ w^{I}_{i,\tau^P_{iT}} \right\}.$$

5.2 Posterior Samplers

5.2.1 Random Coefficients Model

Compared to Subsection 3.1 for the baseline setup, the major change here is to account for cross-sectional heteroskedasticity via another flexible prior on the distribution of σ_i^2. Define $l_i = \log\left(\sigma_i^2 - \underline{\sigma}^2\right)$ where $\underline{\sigma}^2$ is some small positive number. Then, the support of $f_0^{\sigma^2}$ is bounded below by $\underline{\sigma}^2$ and thus satisfies the requirement for the asymptotic convergence of the density forecasts in Proposition

5.12.[17] The log transformation ensures an unbounded support for l_i so that Algorithm 3.1 with Gaussian-mixture DPM prior can be directly employed. Beyond cross-sectional heteroskedasticity, there is a minor alternation due to the (potentially) multivariate λ_i. In this scenario, the component mean μ_k is a vector and component variance Ω_k is a positive definite matrix.

The following algorithm parallels Algorithm 3.1. Both algorithms are based on truncation approximation, which is relatively easy to implement and enjoys good mixing properties. For the slice-retrospective sampler, please refer to Algorithm B.4 in the Appendix.

Denote $D = \{\{D_i\}, D_A\}$ as a shorthand for the data sample used for estimation, where $D_i = c^*_{i,T}$ contains the observed data for individual i, and $D_A = w^A_{0:T}$ is composed of the aggregate regressors with heterogeneous effects. Note that because λ_i and σ_i^2 are independent with respect to each other, their mixture structures are completely separate. As mixture structures of λ_i and l_i are almost identical, I define a generic variable z which can represent either λ or l, and then include z as a superscript to indicate whether a specific parameter belongs to the λ part or the l part. Most of the conditional posteriors are either similar to Algorithm B.4 or standard for posterior sampling (see Appendix B.3), except for the additional term $\left(\sigma_i^2 - \underline{\sigma}^2\right)^{-1}$ in step 4-b, which takes care of the change of variables from $l_i = \log\left(\sigma_i^2 - \underline{\sigma}^2\right)$ to σ_i^2.

Algorithm 5.1. *(General Model: Random Coefficients)*

For each iteration $s = 1, \cdots, n_{sim}$,

1. *Component probabilities: For* $z = \lambda,\ l$,
   (a) *Draw* $\alpha^{z(s)}$ *from a gamma distribution* $p\left(\alpha^{z(s)} \middle| p_{K^z}^{z(s-1)}\right)$.
   (b) *For* $k^z = 1, \cdots, K^z$, *draw* $p_{k^z}^{z(s)}$ *from the truncated stick breaking process* $p\left(\left\{p_{k^z}^{z(s)}\right\} \middle| \alpha^{z(s)}, \left\{n_{k^z}^{z(s-1)}\right\}\right)$.
2. *Component parameters: For* $z = \lambda,\ l$, *for* $k^z = 1, \cdots, K^z$, *draw* $\left(\mu_{k^z}^{z(s)}, \Omega_{k^z}^{z(s)}\right)$ *from a multivariate-normal-inverse-Wishart distribution (or a normal-inverse-gamma distribution if* z *is a scalar)* $p\left(\mu_{k^z}^{z(s)}, \Omega_{k^z}^{z(s)} \middle| \left\{z_i^{(s-1)}\right\}_{i \in J_{k^z}^{z(s-1)}}\right)$.
3. *Component memberships: For* $z = \lambda,\ l$, *for* $i = 1, \cdots N$, *draw* $\gamma_i^{z(s)}$ *from a multinomial distribution* $p\left(\left\{\gamma_i^{z(s)}\right\} \middle| \left\{p_{k^z}^{z(s)}, \mu_{k^z}^{z(s)}, \Omega_{k^z}^{z(s)}\right\},\ z_i^{(s-1)}\right)$.
4. *Individual-specific parameters:*
   (a) *For* $i = 1, \cdots, N$, *draw* $\lambda_i^{(s)}$ *from a multivariate-normal distribution (or a normal distribution if* λ *is a scalar)* $p\left(\lambda_i^{(s)} \middle| \mu_{\gamma_i^\lambda}^{\lambda(s)}, \Omega_{\gamma_i^\lambda}^{\lambda(s)}, \left(\sigma_i^2\right)^{(s-1)}, \beta^{(s-1)}, D_i, D_A\right)$.

[17] Note that only Proposition 5.12 for density forecasts needs a positive lower bound on the distribution of σ_i^2. The propositions for identification and posterior consistency of the estimates are not restricted to but can accommodate such requirement.

(b) *For* $i = 1, \cdots, N$, *draw* $\left(\sigma_i^2\right)^{(s)}$ *via the random-walk Metropolis-Hastings approach*

$$p\left(\left(\sigma_i^2\right)^{(s)} \middle| \mu_{\gamma_i^l}^{l(s)}, \Omega_{\gamma_i^l}^{l(s)}, \lambda_i^{(s)}, \beta^{(s-1)}, D_i, D_A\right)$$
$$\propto \left(\left(\sigma_i^2\right)^{(s)} - \underline{\sigma}^2\right)^{-1} \phi\left(\log\left(\left(\sigma_i^2\right)^{(s)} - \underline{\sigma}^2\right);\ \mu_{\gamma_i^l}^{l(s)}, \Omega_{\gamma_i^l}^{l(s)}\right) \prod_{t=t_{0i}}^{t_{1i}} \phi\left(y_{it};\ \lambda_i^{(s)\prime} w_{i,t-1} + \beta^{(s-1)\prime} x_{i,t-1}, \left(\sigma_i^2\right)^{(s)}\right).$$

5. *Common parameters: Draw* $\beta^{(s)}$ *from a linear regression model* $p\left(\beta^{(s)} \middle| \left\{\lambda_i^{(s)}, \left(\sigma_i^2\right)^{(s)}\right\},\ D\right)$.

5.2.2 Correlated Random Coefficients Model

Regarding conditional density estimation, I impose the MGLR_x prior on both λ_i and l_i. Compared to Algorithm 3.2 for the baseline setup, the algorithm here makes the following changes: (1) generic variable $z = \lambda,\ l$, (2) $\left(\sigma_i^2 - \underline{\sigma}^2\right)^{-1}$ in step 4-b, (3) vector λ_i, and (4) vector conditioning set c_{i0}. The conditioning set c_{i0} is characterized by equation (5.2) for balanced panels or equation (5.3) for unbalanced panels. In practice, it is more computationally efficient to incorporate a subset of c_{i0} or a function of c_{i0} guided by the specific problem at hand.

Algorithm 5.2. *(General Model: Correlated Random Coefficients)*

For each iteration $s = 1, \cdots, n_{sim}$,

1. *Component probabilities: For* $z = \lambda,\ l$,
   (a) *For* $k^z = 1, \cdots, K^z - 1$, *draw* $A_{k^z}^{z(s)}$ *via the random-walk Metropolis-Hastings approach,* $p\left(A_{k^z}^{z(s)} \middle| \boldsymbol{\zeta}_{k^z}^{z(s-1)}, \{c_{i0}\}\right)$ *and then calculate* $\boldsymbol{V}_k^{(s)}$.
   (b) *For* $k^z = 1, \cdots, K^z - 1$, *and* $i = 1, \cdots, N$, *draw* $\xi_{k^z}^{z(s)}(c_{i0})$ *from a truncated normal distribution* $p\left(\xi_{k^z}^{z(s)}(c_{i0}) \middle| \zeta_{k^z}^{z(s-1)}(c_{i0}), \gamma_i^{z(s-1)}\right)$.
   (c) *For* $k^z = 1, \cdots, K^z - 1$, $\boldsymbol{\zeta}_{k^z}^{z(s)}$ *from a multivariate normal distribution* $p\left(\boldsymbol{\zeta}_{k^z}^{z(s)} \middle| \boldsymbol{V}_{k^z}^{z(s)}, \boldsymbol{\xi}_{k^z}^{z(s)}\right)$.
   (d) *For* $k^z = 1, \cdots, K^z - 1$, *and* $i = 1, \cdots, N$, *the component probabilities* $p_{k^z}^{z(s)}(c_{i0})$ *are fully determined by* $\boldsymbol{\zeta}_{k^z}^{z(s)}$.
2. *Component parameters: For* $z = \lambda,\ l$, *for* $k^z = 1, \cdots, K^z$,
   (a) *Draw* $\mu_{k^z}^{z(s)}$ *from a matricvariate-normal distribution (or a multivariate-normal distribution if* z *is a scalar)* $p\left(\mu_{k^z}^{z(s)} \middle| \Omega_{k^z}^{z(s-1)}, \left\{z_i^{(s-1)}, c_{i0}\right\}_{i \in J_{k^z}^{z(s-1)}}\right)$.
   (b) *Draw* $\Omega_{k^z}^{z(s)}$ *from an inverse-Wishart distribution (or an inverse-gamma distribution if* z *is a scalar)* $p\left(\Omega_{k^z}^{z(s)} \middle| \mu_{k^z}^{z(s)}, \left\{z_i^{(s-1)}, c_{i0}\right\}_{i \in J_{k^z}^{z(s-1)}}\right)$.
3. *Component memberships: For* $z = \lambda,\ l$, *for* $i = 1, \cdots N$, *draw* $\gamma_i^{z(s)}$ *from a multinomial distribution* $p\left(\left\{\gamma_i^{z(s)}\right\} \middle| \left\{p_{k^z}^{z(s)}, \mu_{k^z}^{z(s)}, \Omega_{k^z}^{z(s)}\right\},\ z_i^{(s-1)}, c_{i0}\right)$.
4. *Individual-specific parameters:*
   (a) *For* $i = 1, \cdots, N$, *draw* $\lambda_i^{(s)}$ *from a multivariate-normal distribution (or a normal distribution if* λ *is a scalar)* $p\left(\lambda_i^{(s)} \middle| \mu_{\gamma_i^\lambda}^{\lambda(s)}, \Omega_{\gamma_i^\lambda}^{\lambda(s)}, \left(\sigma_i^2\right)^{(s-1)}, \beta^{(s-1)}, D_i, D_A\right)$.

(b) *For* $i = 1, \cdots, N$, *draw* $\left(\sigma_i^2\right)^{(s)}$ *via the random-walk Metropolis-Hastings approach* $p\left(\left(\sigma_i^2\right)^{(s)} \middle| \mu_{\gamma_i^l}^{l(s)}, \Omega_{\gamma_i^l}^{l(s)}, \lambda_i^{(s)}, \beta^{(s-1)}, D_i, D_A\right)$.

5. *Common parameters: Draw* $\beta^{(s)}$ *from a linear regression model* $p\left(\beta^{(s)} \middle| \left\{\lambda_i^{(s)}, \left(\sigma_i^2\right)^{(s)}\right\}, D\right)$.

5.3 Identification

5.3.1 Balanced Panels

Assumption 5.3. *(General Model: Setup)*

1. *Conditional on* $w_{0:T}^A$, $\left\{c_{i0}^*, \lambda_i, \sigma_i^2\right\}$ *are i.i.d. across* i.
2. *For all* t, *conditional on* $\{y_{it}, c_{i,t-1}\}$, x_{it}^{P*} *is independent of* $\left\{\lambda_i, \sigma_i^2\right\}$ *and* β.
3. $\left\{x_{i,0:T}^O, w_{i,0:T}\right\}$ *are independent of* $\left\{\lambda_i, \sigma_i^2\right\}$ *and* β.
4. *Let* $u_{it} = \sigma_i v_{it}$. v_{it} *is i.i.d. across* i *and* t *and independent of* $c_{i,t-1}$.

Remark 5.4. (i) For the random effects case, the first condition can be altered to "$\left\{\lambda_i, \sigma_i^2\right\}$ are independent of c_{i0} and i.i.d. across i".

(ii) For the distribution of the shock u_{it}, a general class of shock distributions can be accommodated by the scale mixture of normals generated from the flexible distribution of σ_i^2 (Kelker, 1970; Fernandez and Steel, 2000). It is possible to allow some additional flexibility in the distribution of u_{it}. For example, the identification argument still holds as long as (1) v_{it} is i.i.d. across i and independent over t, and (2) the distributions of v_{it}, $f_t^v\left(v_{it}\right)$, have known functional forms, such that $\mathbb{E}[v_{it}] = 0$, $\mathbb{V}[v_{it}] = 1$. Nevertheless, as this paper studies panels with short time spans, time-varying shock distribution may not play a significant role. I will keep the normality assumption in the rest of this paper to streamline the arguments.

Assumption 5.5. *(General Model: Identification) For all* i,

1. *The common parameter vector* β *is identifiable.*[18]
2. $w_{i,0:T-1}$ *has full rank* d_w.
3. *Conditioning on* c_{i0}, λ_i *and* σ_i^2 *are independent of each other.*
4. *The characteristic functions for* $\lambda_i|c_{i0}$ *and* $\sigma_i^2|c_{i0}$ *are non-vanishing almost everywhere.*

Proposition 5.6. *(General Model: Identification)*

Under Assumptions 5.3 and 5.5, the common parameters β *and the conditional distribution of individual effects,* $f^{\lambda}(\lambda_i|c_{i0})$ *and* $f^{\sigma^2}(\sigma_i^2|c_{i0})$, *are all identified.*

Please refer to Appendix D.1 for the proof. Assumption 5.3-5.5 and Proposition 5.6 are similar to Assumption 2.1-2.2 and Theorem 2.3 in Liu *et al.* (2016) except for the treatment of heteroskedasticity. First, this paper supports unobserved cross-sectional heteroskedasticity whereas Liu *et al.*

[18] The identification of common parameters in panel data models is standard in the literature. For example, there have been various ways to difference data across t to remove the individual effects λ_i (e.g. orthogonal forward differencing, see Appendix D.1), and we can construct moment conditions based on the transformed data to identify the common parameters β. Here I follow Liu *et al.* (2016) and state a high-level identification assumption.

(2016) incorporate cross-sectional heteroskedasticity as a parametric function of observables. Second, Liu *et al.* (2016) allow for time-varying heteroskedasticity whereas the identification restriction in this paper can only permit time-varying distribution for v_{it} (see Remark 5.4 (ii)) while keeping zero mean and unit variance. However, considering that this paper focuses on the scenarios with short time dimension, lack of time-varying heteroskedasticity would not be a major concern.

5.3.2 Unbalanced Panels

Assumption 5.7. *(Unbalanced Panels) For all i,*

1. c_{i0} *is observed.*
2. x_{iT} *and* w_{iT} *are observed.*
3. *The common parameter vector* β *is identifiable.*
4. $w_{i,(t_{0i}-1):(t_{1i}-1)}$ *has full rank* d_w.

The first condition guarantees the existence of the initial conditioning set for the correlated random coefficients model. In practice, it is not necessary to incorporate all initial values of the predetermined variables and the whole series of the strictly exogenous variables. It is more feasible to only take into account a subset of c_{i0} or a function of c_{i0} that is relevant for the specific analysis. The second condition ensures that the covariates in the forecast equation are available in order to make predictions. The third condition is the same as Assumption 5.5 (1) that makes a high-level assumption on the identification of common parameters. The fourth condition is the unbalanced panel counterpart of Assumption 5.5 (2). It guarantees that the observed chain is long and informative enough to distinguish different aspects of individual effects. Now we can state similar identification results for unbalanced panels.

Proposition 5.8. *(Identification: Unbalanced Panels)*

For unbalanced panels, under Assumptions 5.3, 5.5 (3-4), and 5.7, the common parameter vector β *and the conditional distributions of individual effects,* $f^{\lambda}(\lambda_i|c_{i0})$ *and* $f^{\sigma^2}(\sigma_i^2|c_{i0})$*, are all identified.*

5.4 Asymptotic Properties

In Subsection 5.4.1, I address posterior consistency of f^{σ^2} with unknown individual-specific heteroskedasticity σ_i^2. In Subsection 5.4.2, I proceed with the general setup (5.1) by considering (correlated) random coefficients, adding other strictly exogenous and predetermined covariates into x_{it}, and accounting for unbalanced panels, then the posterior consistency can be obtained with respect to the common parameters vector β and the (conditional) distributions of individual effects, f^{λ} and f^{σ}. In Subsection 5.4.3, I establish the asymptotic properties of the density forecasts.

Let d_z be the dimension of z_{it}, where z is a generic vector of variables which can be either w (observables with heterogeneous effects) or x (observables with homogeneous effects). Then, the space of common parameters $\Theta = \mathbb{R}^{d_x}$, the space of distributions of heterogeneous coefficients $\mathcal{F}^{\lambda}$

is a set of (conditional) densities on $\mathbb{R}^{d_w}$, and the space of distributions of shock sizes $\mathcal{F}^{\sigma^2}$ is a set of (conditional) densities on $\mathbb{R}^+$. The data sample used for estimation is $D = \{\{D_i\}, D_A\}$ defined in Subsection 5.2.1, which constitutes the conditioning set for posterior inference.

5.4.1 Cross-sectional Heteroskedasticity

In many empirical applications, such as the young firm analysis in Section 7, risk may largely vary over the cross-section. Therefore, it is more realistic to address cross-sectional heteroskedasticity, which also contributes considerably to more precise density forecasts. To illustrate the main essence, let us adapt the special case in equation (4.4) to incorporate cross-sectional heteroskedastic shocks while keeping random effects and balanced panels unchanged.

$$y_{it} = \lambda_i + u_{it}, \quad u_{it} \sim N\left(0, \sigma_i^2\right), \tag{5.4}$$

where $\beta = 0$, and λ_i is independent of σ_i^2. Their distributions, $f^{\lambda}(\lambda_i)$ and $f^{\sigma^2}\left(\sigma_i^2\right)$, are unknown, with the true distributions being $f_0^{\lambda}(\lambda_i)$ and $f_0^{\sigma^2}\left(\sigma_i^2\right)$, respectively. Their posteriors are consistently estimated as established in the following proposition.

Proposition 5.9. *(Cross-sectional Heteroskedasticity)*

In setup (5.4) with the random effects version of Assumption 5.3 (1 and 4) and Assumption 5.5 (3-4), if $f_0^{\lambda} \in KL\left(\Pi^{f^{\lambda}}\right)$ *and* $f_0^{\sigma^2} \in KL\left(\Pi^{f^{\sigma^2}}\right)$*, the posterior is weakly consistent at* $\left(f_0^{\lambda},\ f_0^{\sigma^2}\right)$.

Please refer to Appendix D.2 for the complete proof. The KL requirement is again given by the convexity of KL divergence. The intuition of the tests is again to break down the alternatives into two circumstances. First, when a candidate f^{σ^2} and the true $f_0^{\sigma^2}$ are not identical, we can once again rely on orthogonal forward differencing (see Appendix D.1) to distinguish variance distributions. Note that the Fourier transformation (i.e. characteristic functions) is not suitable for disentangling products of random variables, so I resort to the Mellin transform (Galambos and Simonelli, 2004) instead. The second circumstance comes when the variance distributions are close to each other, but f^{λ} is far from f_0^{λ}. Here I apply the argument for Proposition 4.7 with slight adaption.

$f_0^{\lambda} \in KL\left(\Pi^{f^{\lambda}}\right)$ is guaranteed by the sufficient conditions in Lemma 4.8 (or Lemma E.1 for true distribution with heavy tails). Concerning $f_0^{\sigma^2}$, I impose a Gaussian-mixture DPM prior on $l = \log\left(\sigma^2 - \underline{\sigma}^2\right)$, and similar sufficient conditions apply to the distribution of l as well.

5.4.2 General Setup

In this subsection, I generalize the setup to the full panel data model in equation (5.1) with regard to the following three aspects. The proofs are along the same lines of the baseline model plus cross-sectionally heteroskedasticity.

First, in practice, it is more desirable to consider heterogeneous coefficients beyond the individual-specific intercept, which features a vector of λ_i interacting with observed w_{it}. In the young firm

example, different young firms may respond differently to the financial crisis, and R&D activities may benefit the young firms in different magnitudes. A (correlated) random coefficient model can capture such heterogeneities and facilitate predictions.

The uniformly exponentially consistent tests for multivariate λ_i are constructed in a similar way as Proposition 4.7 outlined in the "disentangle skills and shocks" part of Subsection 4.3.1. Note that for each $l = 1, \cdots, d_w$, we can implement orthogonal forward differencing with respect to all other $\{\lambda_{im}\}_{m \neq l}$ and reduce the problem to λ_{il} versus shocks as in equation (4.3). The same logic still holds when we add lagged dependent variables and other predictors. Furthermore, a multi-dimensional version of Lemma 4.8 or Assumption 4.14 guarantees the KL property of multivariate λ_i .

Second, additional strictly exogenous ($x^O_{i,t-1}$) and predetermined ($x^{P*}_{i,t-1}$) predictors help control for other sources of variation and gain more accurate forecasts. We can reproduce the proof of Proposition 4.15 by allowing the conditioning set c_{i0} to include the initial values of the predetermined variables and the whole series of the strictly exogenous variables.

Third, it is constructive to account for unbalanced panels with missing observations, which incorporates more data into the estimation and elicits more information for the prediction. Conditional on the covariates, the common parameters, and the distributions of individual heterogeneities, $y_{i,t+1}$s are cross-sectionally independent, so the posterior consistency argument is still valid in like manner given Assumption 5.7.

Combining above discussions all together, we achieve the posterior consistency result for the general panel data model. The random coefficients model is relatively more straightforward regarding posterior consistency, as the random coefficients setup together with Assumption 5.5 (3) implies that $\left(\lambda_i, \sigma_i^2, c_{i0}\right)$ are independent among one another. The theorem for the random coefficients model is stated as follows.

Proposition 5.10. *(General Model: Random Coefficients)*

Suppose we have:

1. *Assumptions 5.3, 5.5 (3-4), 5.7, and 4.10.*
2. *Lemma 4.8 on λ and l.*
3. $\beta_0 \in \text{supp}\left(\Pi^\beta\right)$.

Then, the posterior is weakly consistent at $\left(\beta_0, f_0^\lambda, f_0^{\sigma^2}\right)$.

For heavy tails in the true unknown distributions, Lemma E.2 generalizes Lemma E.1 to the multivariate scenario, and Proposition E.3 gives a parallel posterior consistency result.

In the world of correlated random coefficients, λ_i is independent of σ_i^2 conditional on c_{i0}. In other words, λ_i and σ_i^2 can potentially depend on the initial condition c_{i0}, and therefore can potentially relate to each other through c_{i0}. For example, a young firm's initial performance may reveal its underlying ability and risk. The following proposition is established for the correlated random coefficients model.

Proposition 5.11. *(General Model: Correlated Random Coefficients)*

Under Assumptions 5.3, 5.5 (3-4), 5.7, 4.10, and 4.14, if $\beta_0 \in supp\left(\Pi^\beta\right)$*, the posterior is weakly consistent at* $\left(\beta_0, f_0^\lambda, f_0^{\sigma^2}\right)$.

Note that Propositions 5.10 and 5.11 are parallel with each other, as the first group of conditions in Assumption 4.14 is the conditional analog of Lemma 4.8 conditions.

5.4.3 Density Forecasts

In the sequel, the next proposition shows convergence of density forecasts in the general model.

Proposition 5.12. *(General Model: Density Forecasts)*

In the general model (5.1), suppose we have:

1. *For the random coefficients model,*
   (a) *conditions in Proposition 5.10,*
   (b) $supp\left(f_0^{\sigma^2}\right)$ *is bounded below by some* $\underline{\sigma}^2 > 0$.
2. *For the correlated random coefficients model,*
   (a) *conditions in Proposition 5.11,*
   (b) $q_0(y_0)$ *is continuous, and there exists* $\underline{q} > 0$ *such that* $|q_0(y_0)| > \underline{q}$ *for all* $y_0 \in \mathcal{C}$,
   (c) $supp\left(f_0^{\sigma^2}\right)$ *is bounded below by some* $\underline{\sigma}^2 > 0$.

Then the density forecasts converge to the oracle predictor in the following two ways:

1. *Convergence of* $f_{i,T+1}^{cond}$ *in weak topology: for any* i *and any* $U_{\epsilon,\Phi}\left(f_{i,T+1}^{oracle}\right)$, *as* $N \to \infty$,

$$\mathbb{P}\left(f_{i,T+1}^{cond} \in U_{\epsilon,\Phi}\left(f_{i,T+1}^{oracle}\right) \middle| y_{1:N,0:T}\right) \to 1, \ a.s.$$

2. *"Pointwise" convergence of* $f_{i,T+1}^{sp}$*: for any* i, *any* y, *and any* $\epsilon > 0$, *as* $N \to \infty$,

$$\left|f_{i,T+1}^{sp}(y) - f_{i,T+1}^{oracle}(y)\right| < \epsilon, \ a.s.$$

The additional requirement that the support of $f_0^{\sigma^2}$ is bounded below ensures that the likelihood would not explode. Then, the proof is in the same vein as the baseline setup.

6 Simulation

In this section, I have conducted extensive Monte Carlo simulation experiments to examine the numerical performance of the proposed semiparametric Bayesian predictor. Subsection 6.1 describes the evaluation criteria for point forecasts and density forecasts. Subsection 6.2 introduces other alternative predictors. Subsection 6.3 considers the baseline setup with random effects. Subsection 6.4 extends to the general setup incorporating cross-sectional heterogeneity and correlated random coefficients.

6.1 Forecast Evaluation Methods

As mentioned in the model setup in Subsection 2.1, this paper focuses on one-step-ahead forecasts, but a similar framework can be applied to multi-period-ahead forecasts. The forecasting performance is evaluated along both the point and density forecast dimensions, with particular attention to the latter.

Point forecasts are evaluated via the Mean Square Error (MSE), which corresponds to the quadratic loss function. Let $\hat{y}_{i,T+1}$ denote the forecast made by the model,

$$\hat{y}_{i,T+1} = \hat{\beta}' x_{iT} + \hat{\lambda}_i' w_{iT},$$

where $\hat{\lambda}_i$ and $\hat{\beta}$ stand for the estimated parameter values. Then, the forecast error is defined as

$$\hat{e}_{i,T+1} = y_{i,T+1} - \hat{y}_{i,T+1},$$

with $y_{i,T+1}$ being the realized value at time $T+1$. The formula for the MSE is provided in the following equation,

$$MSE = \frac{1}{N} \sum_i \hat{e}_{i,T+1}^2.$$

The Diebold and Mariano (1995) test is further implemented to assess whether or not the difference in the MSE is significant.

The accuracy of the density forecasts is measured by the log predictive score (LPS) as suggested in Geweke and Amisano (2010),

$$LPS = \frac{1}{N} \sum_i \log \hat{p}\left(y_{i,T+1}|D\right),$$

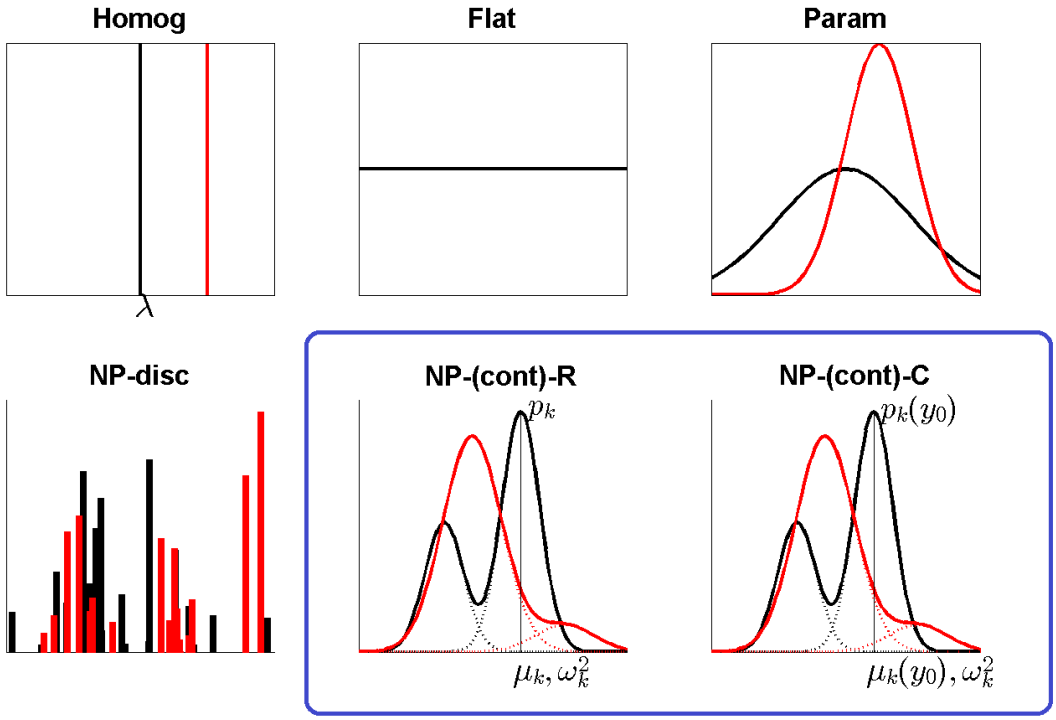
where $y_{i,T+1}$ is the realization at $T+1$, and $\hat{p}\left(y_{i,T+1}|D\right)$ represents the predictive likelihood with respect to the estimated model conditional on the observed data D. In addition, $\exp\left(LPS_A - LPS_B\right)$ gives the odds of the future realizations based on predictor A versus predictor B. I also perform the Amisano and Giacomini (2007) test to examine the significance in the LPS difference.

6.2 Alternative Predictors

In the simulation experiments, I compare the proposed semiparametric Bayesian predictor with alternatives. Different predictors are characterized by different priors. As these priors are distributions over distributions, Figure 6.1 plots two draws from each prior – one in red and the other in black.

The homogeneous prior (Homog) implies an extreme kind of pooling, which assumes that all firms share the same level of skill λ^*. It can be viewed as a Bayesian counterpart of the pooled OLS estimator. Because λ^* is unknown beforehand, the corresponding subgraph plots two vertical lines representing two degenerate distributions with different locations. More rigorously, this prior

Figure 6.1: Alternative Predictors



The black and red lines represent two draws from each prior.

is defined as $\lambda_i \sim \delta_{\lambda^*}$, where δ_{λ^*} is the Dirac delta function representing a degenerate distribution $\mathbb{P}(\lambda_i = \lambda^*) = 1$. The unknown λ^* becomes another common parameter, similar to β, so I adopt a multivariate-normal-inverse-gamma prior on $\left([\beta,\ \lambda^*]',\ \sigma^2\right)$.

The flat prior (Flat) is specified as $p(\lambda_i) \propto 1$, an uninformative prior with the posterior mode being the MLE estimate. Roughly speaking, given the common parameters, there is no pooling from the cross-section, so we learn firm i's skill λ_i only using its own history.

The parametric prior (Param) pools the information from the cross-section via a parametric skill distribution, such as a Gaussian distribution with unknown mean and variance. The corresponding subgraph contains two curves with different means and variances. More explicitly, we have $\lambda_i \sim N\left(\mu_i, \omega_i^2\right)$ where a normal-inverse-gamma hyperprior is further imposed on $\left(\mu_i, \omega_i^2\right)$. This prior can be thought of as a limit case of the DPM prior when the scale parameter $\alpha \to \infty$, so there is only one component, and $\left(\mu_i, \omega_i^2\right)$ are directly drawn from the base distribution G_0. The choice of hyperprior follows the suggestion by Basu and Chib (2003) to match the Gaussian model with the DPM model such that "the predictive (or marginal) distribution of a single observation is identical under the two models" (pp. 226-227).

The nonparametric discrete prior (NP-disc) is modeled by a DP where λ_i follows a flexible nonparametric distribution but on a discrete support. This paper focuses on continuous f, which may be more sensible for the skill of young firms as well as other similar empirical studies. In this sense, it is helpful to check with the "NP-disc" predictor to examine how much can be gained or lost from the continuity assumption and from the additional layer of mixture.

In addition, "NP-R" denotes the proposed nonparametric prior for random effects/coefficients models, and "NP-C" for correlated random effects/coefficients models. Both of them are flexible priors on continuous distributions while "NP-C" allows λ_i to depend on the initial condition of the firms.

The nonparametric predictors would reduce the estimation bias due to their flexibility while increasing the estimation variance due to their complexity. In ex-ante, it is not transparent which predictor performs better – the parsimonious parametric ones or the flexible nonparametric ones. Therefore, it is worthwhile to implement the Monte Carlo experiments and assess which predictor produces more accurate forecasts under which circumstances.

6.3 Baseline Model

Let us first consider the baseline model with random effects. The specifications are summarized in Table 6.1.

β_0 is set to be 0.8 as economic data usually exhibit some degree of persistence. σ_0^2 equals 1/4, so the rough magnitude of signal-noise ratio is $\sigma_0^2/\mathbb{V}(\lambda_i) = 1/4$. The initial conditions y_{i0} is drawn from a truncated normal distribution where I take the standard normal as the base distribution and truncate it at $|y_{i0}| < 5$. This truncation setup complies with Assumption 4.10 such that y_{i0} is

Table 6.1: Simulation Setup: Baseline Model

(a) Dynamic Panel Data Model

Law of motion	$y_{it} = \beta y_{i,t-1} + \lambda_i + u_{it},\ u_{it} \sim N\left(0, \sigma^2\right)$
Common parameters	$\beta_0 = 0.8,\ \sigma_0^2 = 1$
Initial conditions	$y_{i0} \sim TN\left(0, 1, -5, 5\right)$
Sample size	$N = 1000,\ T = 6$

(b) Random Effects

Degenerate	$\lambda_i = 0$
Skewed	$\lambda_i \sim \frac{1}{9}N\left(2, \frac{1}{2}\right) + \frac{8}{9}N\left(-\frac{1}{4}, \frac{1}{2}\right)$
Fat tail	$\lambda_i \sim \frac{1}{5}N\left(0, 4\right) + \frac{4}{5}N\left(0, \frac{1}{4}\right)$
Bimodal	$\lambda_i \sim 0.35N\left(0, 1\right) + 0.65N\left(10, 1\right)$, normalized to $Var\left(\lambda_i\right) = 1$

compactly supported. Choices of $N = 1000$ and $T = 6$ are comparable with the young firm dynamics application.

There are four experiments with different true distributions of λ_i, $f_0\left(\cdot\right)$. As this subsection focuses on the simplest baseline model with random effects, λ_i is independent of y_{i0} in all these four experiments. The first experiment features a degenerate λ_i distribution, where all firms enjoy the same skill level. Note that it does not satisfy the first condition in Lemma 4.8, which requires the true λ_i distribution to be continuous. The purpose of this distribution is to learn how bad things can go under the misspecification that the true λ_i distribution is completely off the prior support. The second and third experiments are based on skewed and fat tail distributions with the functional forms being borrowed from Monte Carlo design 2 in Liu *et al.* (2016). These two specifications reflect more realistic scenarios in empirical studies. The last experiment portrays a bimodal distribution with asymmetric weights on the two components.

I simulated 1,000 panel datasets for each setup and report the average statistics of these 1,000 repetitions. Forecasting performance, especially the relative rankings and magnitudes, is highly stable across repetitions. In each repetition, I generated 40,000 MCMC draws with the first 20,000 being discarded as burn-in. Based on graphical and statistical tests, the MCMC draws seem to converge to a stationary distribution. Both the Brook-Draper diagnostic and the Raftery-Lewis diagnostic yield desirable MCMC accuracy. For trace plots, prior/posterior distributions, rolling means, and autocorrelation graphs of β, σ^2, α, and λ_1, please refer to Figures F.1 to F.4.

Table 6.2 shows the forecasting comparison among alternative predictors. The point forecasts are evaluated by MSE together with the Diebold and Mariano (1995) test. The performance of the density forecasts is assessed by the LPS and the Amisano and Giacomini (2007) test. For the oracle predictor, the table reports the exact values of MSE and LPS (multiplied by the cross-sectional dimension N). For other predictors, the table reports the percentage deviations from the oracle

Table 6.2: Forecast Evaluation: Baseline Model

	Degenerate		Skewed		Fat Tail		Bimodal	
	MSE	LPS*N	MSE	LPS*N	MSE	LPS*N	MSE	LPS*N
Oracle	0.25	-725	0.29	-798	0.29	-804	0.27	-766
NP-R	0.8%	-4	**0.04%**	**-0.3**	**0.08%**	**-1**	**1.2%**	**-6**
Homog	**0.03%*****	**-0.2*****	32%***	-193***	29%***	-187***	126%***	-424***
Flat	21%***	-102***	1.4%***	-7***	0.3%***	-2***	8%***	-38***
Param	0.8%	-4	0.3%***	-1***	0.1%***	-1.5***	7%***	-34***
NP-disc	**0.03%*****	**-0.2*****	31%***	-206***	29%***	-205***	7%***	-40***

MSE and difference with respect to the oracle LPS*N. The tests are conducted with respect to NP-R, with significance levels indicated by *: 10%, **: 5%, and ***: 1%. The entries in bold indicate the best feasible predictor in each column.

For each experiment, point forecasts and density forecasts share comparable rankings. When the λ_i distribution is degenerate, "Homog" and "NP-disc" are the best, as expected. They are followed by "NP-R" and "Param", and "Flat" is considerably worse. When the λ_i distribution is non-degenerate, there is a substantial gain in both point forecasts and density forecasts from employing the "NP-R" predictor. In the bimodal case, the "NP-R" predictor far exceeds all other competitors. In the skewed and fat tailed cases, the "Flat" and "Param" predictors are second best, yet still significantly inferior to "NP-R". The "Homog" and "NP-disc" predictors yield the poorest forecasts, which suggests that their discrete supports are not able to approximate the continuous λ_i distribution, and even the nonparametric DP prior with countably infinite support ("NP-disc") is far from enough.

Therefore, when researchers believe that the underlying λ_i distribution is indeed discrete, the DP prior ("NP-disc") is a more sensible choice; on the other hand, when the underlying λ_i distribution is actually continuous, the DPM prior (or the MGLR_x prior later for the correlated random effects model) promotes better forecasts. In the empirical application to young firm dynamics, it would be more reasonable to assume continuous distributions of individual heterogeneities in levels, reactions to R&D, and shock sizes, and results show that the continuous nonparametric prior outperforms the discrete DP prior in terms of density forecasts (see Table 7.3).

To investigate why we obtain better forecasts, Figure 6.2 demonstrates the posterior distribution of the λ_i distribution (i.e. a distribution over distributions) for experiments "Skewed", "Fat Tail", and "Bimodal". In each case, the subgraphs are constructed from the estimation results of one of the 1,000 repetitions, with the left subgraph given by the "Param" estimator and the right one by "NP-R". In each subgraph, the black solid line represents the true λ_i distribution, f_0. The blue bands show the posterior distribution of f, $\Pi\left(f \mid y_{1:N,0:T}\right)$.

For the skewed λ_i distribution, the "NP-R" estimator better tracks the peak on the left and the tail on the right. For the λ_i distribution with fat tails, the "NP-R" estimator accommodates the slowly decaying tails, but is still not able to fully mimic the spiking peak. For the bimodal

λ_i distribution, it is not surprising that the "NP-R" estimator captures the M-shape fairly nicely. In summary, the nonparametric prior flexibly approximates a vast set of distributions, which helps provide more precise estimates of the underlying λ_i distributions and consequently more accurate density forecasts. This observation confirms the connection between skill distribution estimation and density forecasts as stated in Propositions 4.11 and 4.16.

I have also considered various robustness checks. In terms of the setup, I have tried different cross-sectional dimensions $N = 100,\ 500,\ 1000,\ 10^5$, different time spans $T = 6,\ 10,\ 20,\ 50$, different persistences $\beta = 0.2,\ 0.5,\ 0.8,\ 0.95$, different sizes of the i.i.d. shocks $\sigma^2 = 1/4$ and 1, which govern the signal-to-noise ratio, and different underlying λ_i distributions including standard normal. In general, the "NP-R" predictor is the overall best for density forecasts except when the true λ_i comes from a degenerate distribution or a normal distribution. In the latter case, the parsimonious "Param" prior coincides with the underlying λ_i distribution and is not surprisingly but only marginally better than the "NP-R" predictor. Roughly speaking, in the context of young firm dynamics, the superiority of the "NP-R" predictor is more prominent when the time series for a specific firm i is not informative enough to reveal its skill but the whole panel can recover the skill distribution and hence firm i's uncertainty due to heterogenous skill. That is, "NP-R" works the better than the alternatives when N is not too small, T is not too long, σ^2 is not too large, and the λ_i distribution is relatively non-Gaussian. For instance, as the cross-sectional dimension N increases, the blue band in Figure 6.2 gets closer to the true f_0 and eventually completely overlaps it (see Figure F.5), which resonates the posterior consistency statement.

In terms of estimators, I have also constructed the posterior sampler for more sophisticated priors, such as the Pitman-Yor process which allows power law tail for clustering behaviors, as well as DPM with skew normal components which better accommodates asymmetric data generating process. They provide some improvement in the corresponding situations, but call for extra computation efforts.

6.4 General Model

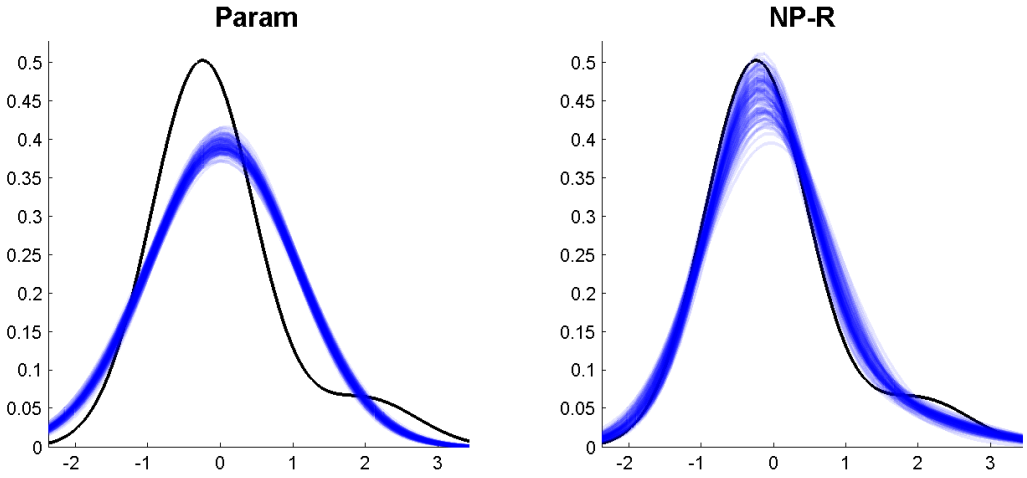
The general model accounts for three key features: (i) multidimensional individual heterogeneity, (ii) cross-sectional heteroskedasticity, and (iii) correlated random coefficients. The exact specification is characterized in Table 6.3.

In terms of multidimensional individual heterogeneity, now λ_i is a 3-by-1 vector, and the corresponding covariates are composed of the level, time-specific $w_{t-1}^{(2)}$, and individual-time-specific $w_{i,t-1}^{(3)}$.

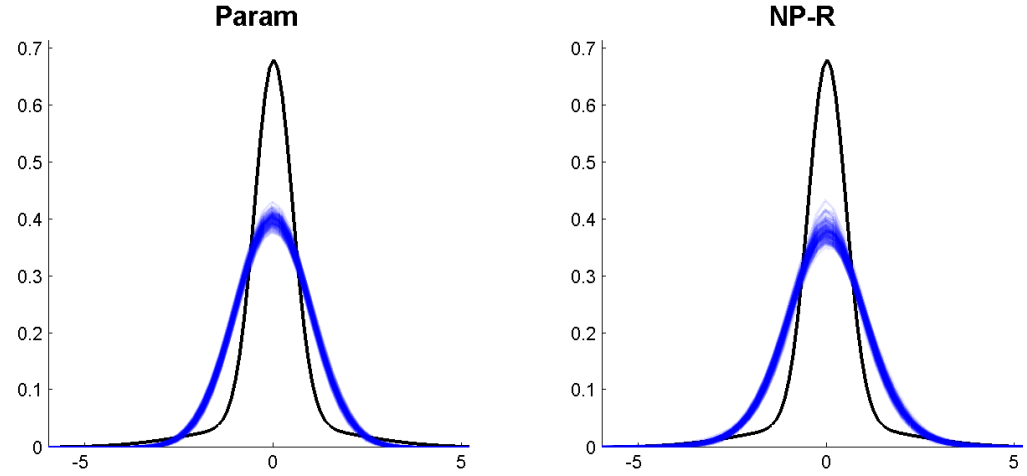
In terms of correlated random coefficients, I adopt the conditional distribution following Dunson and Park (2008) and Norets and Pelenis (2014). They regard it as a challenging problem because such conditional distribution exhibits rapid changes in its shape which considerably restricts local sample size. The original conditional distribution in their papers is one-dimensional, and I expand it

Figure 6.2: f_0 vs $\Pi(f \mid y_{1:N,0:T})$: Baseline Model

(a) Skewed



(b) Fat Tail



(c) Bimodal

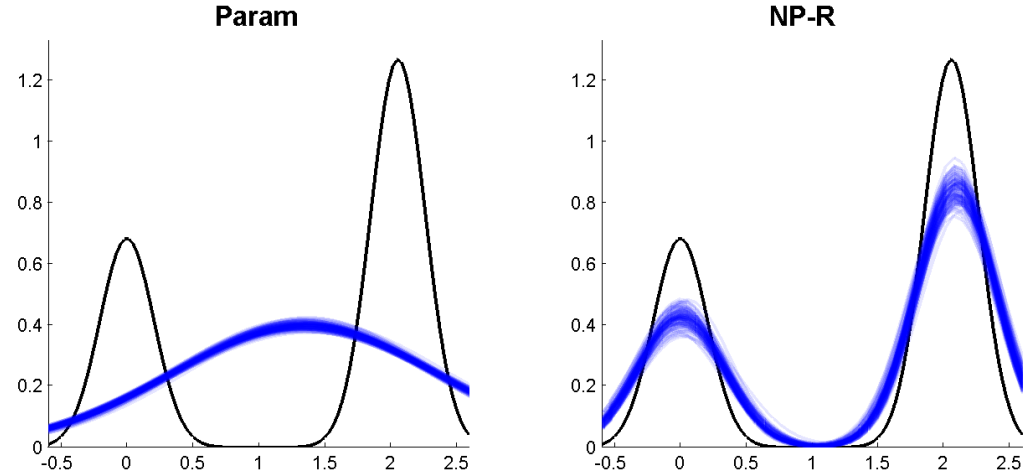


Table 6.3: Simulation Setup: General Model

Law of motion	$y_{it} = \beta y_{i,t-1} + \lambda_i' w_{i,t-1} + u_{it},\ u_{it} \sim N\left(0, \sigma_i^2\right)$
Covariates	$w_{i,t-1} = [1,\ w_{t-1}^{(2)},\ w_{i,t-1}^{(3)}]'$, where $w_{t-1}^{(2)} \sim N(0,1)$ and $w_{i,t-1}^{(3)} \sim \text{Ga}(1,1)$
Common parameters	$\beta_0 = 0.8$
Initial conditions	$y_{i0} \sim U(0,1)$
Correlated random coefficients	$\lambda_i \vert y_{i0} \sim e^{-2y_{i0}} N\left(y_{i0} v,\ 0.1^2 vv'\right) + \left(1 - e^{-2y_{i0}}\right) N\left(y_{i0}^4 v,\ 0.2^2 vv'\right)$, where $v = [1,\ 2,\ -1]'$
Cross-sectional heteroskedasticity	$\sigma_i^2 \vert y_{i0} \sim 0.454 \left(y_{i0} + 0.5\right)^2 \cdot (\text{IG}(51, 40) + 0.2)$
Sample size	$N = 1000,\ T = 6$

Table 6.4: Prior Structures

Predictor		λ_i prior	l_i prior
Heterosk	NP-C	MGLR_x	MGLR_x
Homog		Point mass	Point mass
Homosk	NP-C	MGLR_x	Point mass
Heterosk	Flat	Uninformative	Uninformative
	Param	N	IG
	NP-disc	DP	DP
	NP-R	DPM	DPM

to accommodate the three-dimensional λ_i via a linear transformation of the original. In Figure 6.3 panel (a), the left subgraph presents the joint distribution of λ_{i1} and y_{i0}, where λ_{i1} is the coefficient on $w_{i,t-1}^{(1)} = 1$ and can be interpreted as the heterogeneous intercept. It shows that the shape of the joint distribution is fairly complex, containing many local peaks and valleys. The right subgraph shows the conditional distribution of λ_{i1} given $y_{i0} = 0.25,\ 0.5, 0.75$. We can see that the conditional distribution is involved as well and evolves with the conditioning variable y_{i0}.

In addition, I also let the cross-sectional heteroskedasticity interact with the initial conditions, and the functional form is modified from Pelenis (2014) case 2. The modification guarantees the continuity of σ_i^2 distribution, bounds it above zero (see conditions for Propositions 5.10-5.12), and ensures that the signal-to-noise ratio is not far from 1. Their joint and conditional distributions are depicted in Figure 6.3 panel (b).

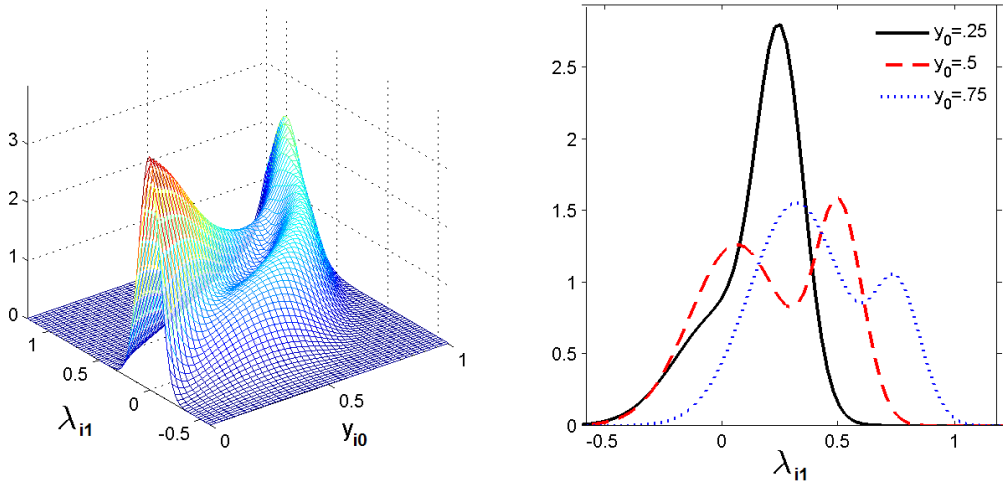
The rest of the setup is the same as the baseline scenario in the previous subsection.

Due to cross-sectional heteroskedasticity and correlated random coefficients, the prior structures become more complicated. Table 6.4 describes the prior setups of λ_i and l_i, with the predictor labels being consistent with the definitions in Subsection 6.2. Note that I further add the "Homosk-NP-C" predictor in order to examine whether it is practically relevant to model heteroskedasticity.

Table 6.5 assesses the forecasting performance of these predictors. Considering point forecasts, from the best to the worst, the ranking is "Heterosk-NP-R", "Heterosk-Param", "Heterosk-NP-disc",

Figure 6.3: DGP: General Model

(a) $p\left(\lambda_{i1}|y_{i0}\right)$



(b) $p\left(\sigma_i^2|y_{i0}\right)$

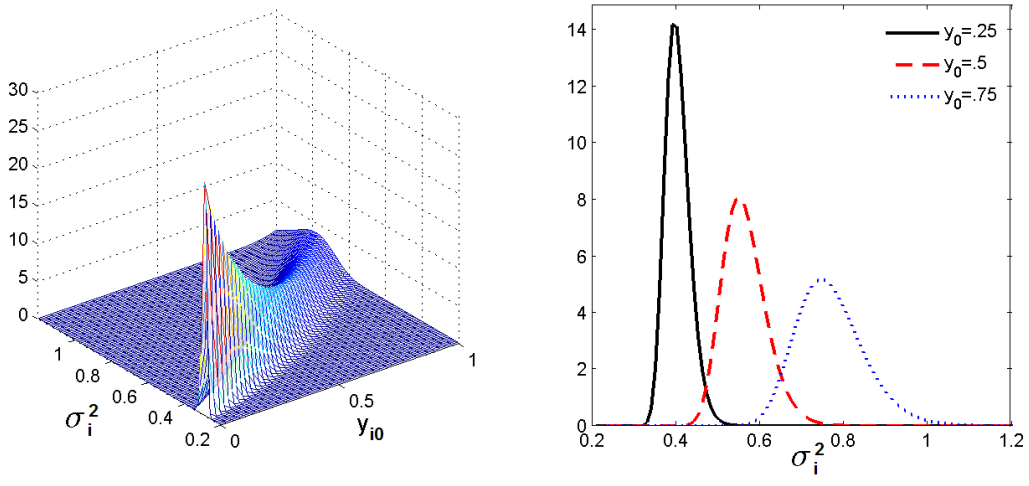


Table 6.5: Forecast Evaluation: General Model

		MSE	LPS*N
Oracle		0.70	-1150
Heterosk	NP-C	13.68%	**-74**
Homog		89.28%***	-503***
Homosk	NP-C	20.84%***	-161***
Heterosk	Flat	151.60%***	-515***
	Param	11.30%*	-139***
	NP-disc	13.08%	-150***
	NP-R	**11.25%***	-93*

The point forecasts are evaluated by the Mean Square Error (MSE) together with the Diebold and Mariano (1995) test. The performance of the density forecasts is assessed by the log predictive score (LPS) and the Amisano and Giacomini (2007) test. For the oracle predictor, the table reports the exact values of MSE and LPS. For other predictors, the table reports the percentage deviations from the benchmark MSE and difference with respect to the benchmark LPS. The tests are conducted with respect to Heterosk-NP-C, with significance levels indicated by *: 10%, **: 5%, ***: 1%. The entries in bold indicate the best feasible predictor in each column.

"Heterosk-NP-C", "Homosk-NP-C", "Homog", and "Heterosk-Flat". The first two constitute the first tier, the next two can be viewed as the second tier, the next one is the third tier, and the last two are markedly inferior. It is not surprising that more parsimonious estimators outperform "Heterosk-NP-C" in terms of point forecasts, though "Heterosk-NP-C" is correctly specified while the parsimonious ones are not.

Nevertheless, the focus of this paper is density forecasting, where "Heterosk-NP-C" becomes the most accurate density predictor. Several lessons can be inferred from a more detailed comparison among predictors. First, based on the comparison between "Heterosk-NP-C" and "Homog"/"Homosk-NP-C", it is important to account for individual effects in both coefficients λ_is and shock sizes σ_i^2s. Second, comparing "Heterosk-NP-C" with "Heterosk-Flat"/"Heterosk-Param", we see that the flexible nonparametric prior plays a significant role in enhancing density forecasts. Third, the difference between "Heterosk-NP-C" and "Heterosk-NP-disc" indicates that the discrete prior performs less satisfactorily when the underlying individual heterogeneity is continuous. Last, "Heterosk-NP-R" is less favorable than "Heterosk-NP-C", which necessitates a careful modeling of the correlated random coefficient structure.

7 Empirical Application: Young Firm Dynamics

7.1 Background and Data

To see how the proposed predictor works in real world analysis, I applied it to provide density forecasts of young firm performance. Studies have documented that young firm performance is

affected by R&D, recession, etc. and that different firms may react differently to these factors (Akcigit and Kerr, 2010; Robb and Seamans, 2014; Zarutskie and Yang, 2015). In this empirical application, I examine these channels from a density forecasting perspective.

To analyze firm dynamics, traditional cross-sectional data are not sufficient whereas panel data are more suitable as they track the firms over time. In particular, it is desirable to work with a dataset that contains sufficient information on early firm financing[19] and innovation, and spreads over the recent recession. The restricted-access Kauffman Firm Survey (KFS) is the ideal candidate for such purpose, as it offers the largest panel of startups (4,928 firms founded in 2004, nationally representative sample) and longest time span (2004-2011, one baseline survey and seven follow-up annual surveys), together with detailed information on young firms. For further description of the survey design, please refer to Robb *et al.* (2009).[20]

7.2 Model Specification

I consider the general model with multidimensional individual heterogeneity in λ_i and cross-sectional heteroskedasticity in σ_i^2. Following the firm dynamics literature, such as Akcigit and Kerr (2010) and Zarutskie and Yang (2015), firm performance is measured by employment. From an economic point of view, young firms make a significant contribution to employment and job creation (Haltiwanger *et al.*, 2012), and their struggle during the recent recession may partly account for the recent jobless recovery. Specifically, here y_{it} is chosen to be the log of employment denoted as $\log \text{emp}_{it}$. I adopt the log of employment instead of employment growth rate since the latter significantly reduces the cross-sectional sample size due to the rank requirment for unbalanced panels. It is preferable to work with larger N according to the theoretical argument.

For the key variables with potential heterogeneous effects ($w_{i,t-1}$), I compare the forecasting performance of the following three setups:[21]

(i) $w_{i,t-1} = 1$, which specifies the baseline model with λ_i being the individual-specific intercept.

(ii) $w_{i,t-1} = [1,\ \text{rec}_{t-1}]'$. rec_t is an aggregate dummy variable indicating the recent recession. It is equal to 1 for 2008 and 2009, and is equal to 0 for other periods.

(iii) $w_{i,t-1} = [1,\ \text{R\&D}_{i,t-1}]'$. $\text{R\&D}_{it}$ is given by the ratio of a firm's R&D employment over its total employment, considering that R&D employment has more complete observations compared to other innovation intensity gauges.[22]

[19] In the current version of the empirical exercises, firm financing variables (e.g. capital structure) are not included as regressors because they overly restrict the cross-sectional dimension, but I intend to include them in future work in which I will explicitly model firm exit and thus allow for a larger cross-section.

[20] Here I do not impose weights on firms as the purpose of the current study is forecasting individual firm performance. Further extensions can easily incorporate weights into the estimation procedure.

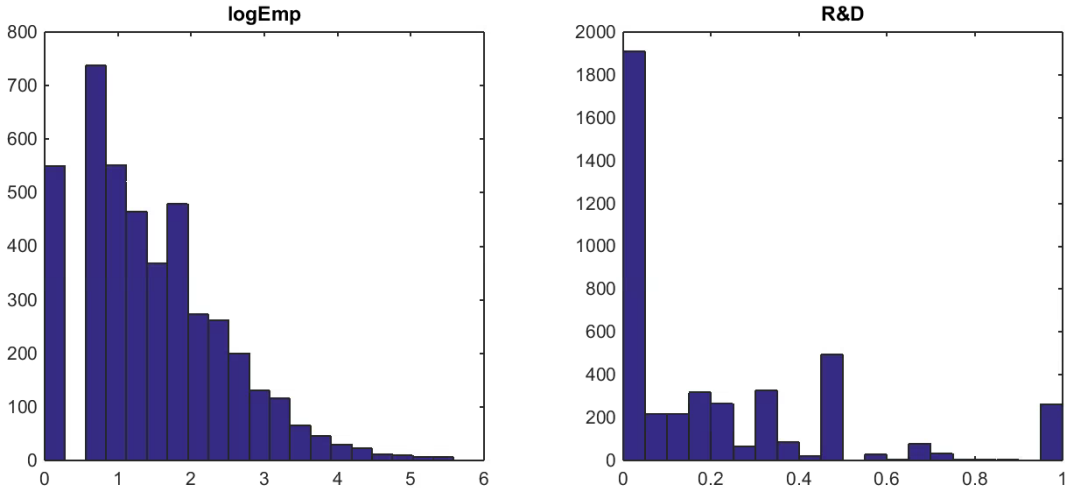
[21] I do not jointly incorporate recession and R&D because such specification largely restricts the cross-sectional sample size due to the rank requirment for unbalanced panels.

[22] I have also explored other measures of firm performance (e.g. the log of revenue) and innovation activities (e.g. a binary variable on whether the firm spends any money on R&D, numbers of intellectual properties–patents, copyrights, or trademarks–owned or licensed by the firm). The estimated AR(1) coefficients and relative rankings of

Table 7.1: Descriptive Statistics for Observable

	10%	mean	med	90%	std	skew	kurt
log emp	0.41	1.44	1.34	2.63	0.86	0.82	3.58
R&D	0.05	0.22	0.17	0.49	0.18	1.21	4.25

Figure 7.1: Histograms for Observables



The panel used for estimation spans 2004 to 2010 with time dimension $T = 6$.[23] The data for 2011 is reserved for pseudo out-of-sample forecast evaluation. Sample selection is performed as follows:

(i) For any (i, t) combination where R&D employment is greater than the total employment, there is an incompatibility issue, so I set $\text{R\&D}_{it} = NA$, which only affects 0.68% of the observations.

(ii) I only keep firms with long enough observations according to Assumption 5.7, which ensures identification in unbalanced panels. This results in cross-sectional dimension $N = 859$ for the baseline specification, $N = 794$ with recession, and $N = 677$ with R&D.

(iii) In order to compare forecasting performance across different setups, the sample is further restricted so that all three setups share exactly the same set of firms.

After all these data cleaning steps, we are left with $N = 654$ firms. The proportion of missing values are $(\#\text{missing obs}) / (NT) = 6.27\%$. The descriptive statistics for $\log \text{emp}_{it}$ and $\text{R\&D}_{it}$ are summarized in Table 7.1, and the corresponding histograms are plotted in Figure 7.1, where both distributions are right skewed and may have more than one peak. Therefore, we anticipate that the proposed predictors with nonparametric priors would perform well in this scenario.

density forecasts are generally robust across measures.

[23] Note that the estimation sample starts from period 0 (i.e. 2004) and ends at period T (i.e. 2010) with $T + 1 = 7$ periods in total.

Table 7.2: Common Parameter β

		Baseline		Recession		R&D	
		mean	std	mean	std	mean	std
Heterosk	NP-C/R	0.48	0.01	0.46	0.02	0.52	0.01
Homog		0.85	0.02	0.85	0.02	0.89	0.02
Homosk	NP-C	0.37	0.02	0.88	0.02	0.51	0.03
Heterosk	Flat	0.19	0.02	0.25	0.00	0.50	0.00
	Param	0.48	0.03	0.26	0.03	0.56	0.03
	NP-disc	0.55	0.02	0.79	0.02	0.84	0.04
	NP-R	0.47	0.03	0.30	0.03	0.74	0.04
	NP-C	0.38	0.02	0.40	0.06	0.53	0.01

7.3 Results

The alternative priors are similar to those in the Monte Carlo simulation except for one additional prior, "Heterosk-NP-C/R", which assumes that λ_i is correlated with y_{i0} while σ_i^2 is not, by imposing an MGLR_x prior on λ_i and a DPM prior on $l_i = \log\left(\sigma_i^2 - \underline{\sigma}^2\right)$. It is possible to craft other priors according to the specific heterogeneity structure of the empirical problem at hand. For example, let λ_{i1} correlate with y_{i0} while setting λ_{i2} independent of y_{i0}. I will leave this to future exploration. The conditioning set is chosen to be standardized y_{i0}. The standardization ensures numerical stability in practice, as the conditioning variables enter exponentially into the covariance function for the Gaussian process.

Table 7.2 characterizes the posterior estimates of the common parameter β. In most of the cases except for "Homog" and "NP-disc", the posterior means are around $0.4 \sim 0.5$, which suggests that the young firm performance exhibits some degree of persistency, but not remarkably strong, which is reasonable as young firms generally experience more uncertainty. For "Homog" and "NP-disc", their posterior means of β are much larger. This may arise from the fact that homogeneous or discrete λ_i structure is not able to capture all individual effects, so these estimators may attribute the remaining individual effects to persistence and thus overestimate β. "NP-R" also gives large estimate of β. The reason is similar – if the true data generating process is correlated random effects/coefficients, the random effects/coefficients model would miss the effects of the initial condition and misinterpret them as the persistence of the system. In all scenarios, the posterior standard deviations are relatively small, which indicates that the posterior distributions are very tight.[24]

Table 7.3 compares the forecasting performance of the predictors across different model setups. The "Heterosk-NP-C/R" predictor is chosen to be the benchmark for all comparisons. For the benchmark predictor, the table reports the exact values of MSE and LPS (multiplied by the cross-

[24]Comparing with the literature, the closest one is Zarutskie and Yang (2015) using usual panel data methods, where the estimated persistence of log employment is 0.824 and 0.816 without firm fixed effects (Table 2) and 0.228 with firm fixed effects (Table 4).

sectional dimension N). For other predictors, the table reports the percentage deviations from the benchmark MSE and difference with respect to the benchmark LPS*N.

In terms of point forecasts, most of the estimators are comparable according to MSE, with only "Flat" performing poorly in all three setups. Intuitively, shrinkage in general leads to better forecasting performance, especially for point forecasts, whereas the "Flat" prior does not introduce any shrinkage to individual effects $\left(\lambda_i, \sigma_i^2\right)$. Conditional on the common parameter β, the "Flat" estimator of $\left(\lambda_i, \sigma_i^2\right)$ is a Bayesian analog of individual-specific MLE/OLS that utilizes only the individual-specific observations, which is inadmissible under fixed T (Robbins, 1956; James and Stein, 1961; Efron, 2012).

For density forecasts measured by LPS, the overall best is the "Heterosk-NP-C/R" predictor in the R&D setup. Comparing setups, the one with recession yields the worst density forecasts (and point forecasts as well), so the recession dummy does not contribute much to forecasting and may even incur overfitting.

Comparing across predictors for the baseline and R&D setups, the main message is similar to the Monte Carlo simulation of the general model in Subsection 6.4. In summary, it is crucial to account for individual effects in both coefficients λ_is and shock sizes σ_i^2s through a flexible nonparametric prior that acknowledges continuity and correlated random effects/coefficients when the underlying individual heterogeneity is likely to possess these features.[25] Note that now both "NP-R" and "NP-C" are inferior to "NP-C/R" where the distribution of λ_i depends on the initial conditions but the distribution of σ_i^2 does not.[26]

Figure 7.2 provides the histograms of the probability integral transformation (PIT) in the R&D setup. While LPS characterizes the relative ranks of predictors, PIT supplements LPS and can be viewed as an absolute evaluation on how good the density forecasts coincide with the true (unobserved) conditional forecasting distributions with respect to the current information set. In this sense, under the null hypothesis that the density forecasts coincide with the truth, the probability integral transforms are i.i.d. $U(0,1)$ and the histogram is close to a flat line. For details of PIT, please refer to Diebold *et al.* (1998). In each subgraph, the two red lines indicate the confidence interval. We can see that, in "NP-C/R", "NP-C" and "Flat", the histogram bars are mostly within the confidence band, while other predictors yield apparent inverse-U shapes. The reason might be that the other predictors do not take correlated random coefficients into account but instead attributes the subtlety of correlated random coefficients to the estimated variance, which leads to more diffused predictive distributions.[27]

Figure 7.3 shows the predictive distributions of 10 randomly selected firms in the R&D setup. In

[25] Intuitively, in the R&D setup, the odds given by the exponential of the difference in LPS indicate that the future realizations are on average 12% more likely in "Heterosk-NP-C/R" versus "Homog", 60% more likely in "Heterosk-NP-C/R" versus "Heterosk-Flat", etc.

[26] This result cannot be directly compared to the Gibrat's law literature (Lee *et al.*, 1998; Santarelli *et al.*, 2006), as the dependent variable here is the log of employment instead of employment growth.

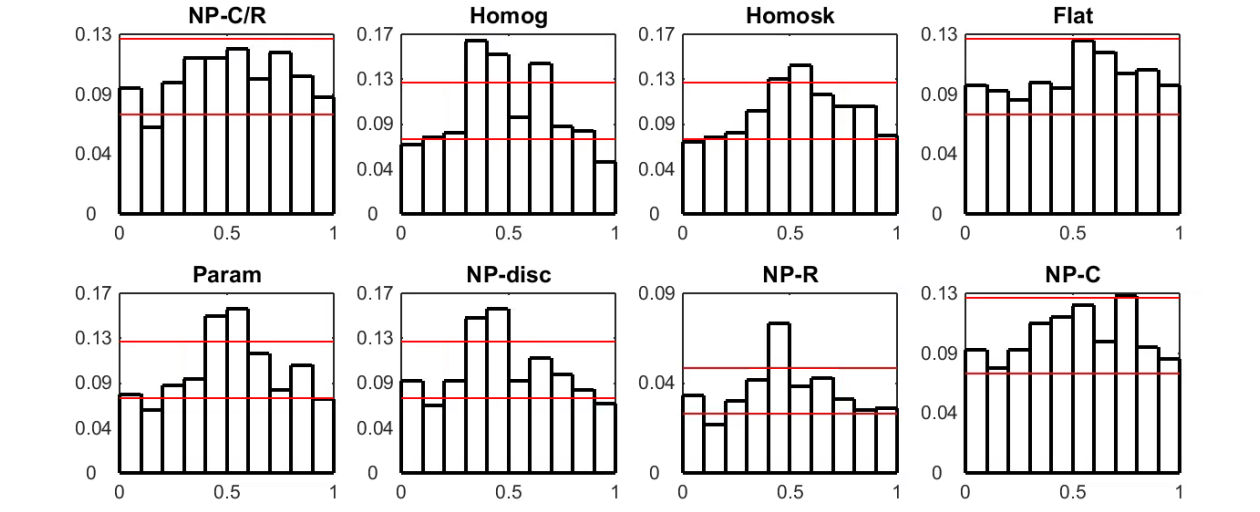
[27] In future revisions, I plan to implement the formal PIT tests proposed in Amisano and Geweke (2016).

Table 7.3: Forecast Evaluation: Young Firm Dynamics

		Baseline		Recession		R&D	
		MSE	LPS*N	MSE	LPS*N	MSE	LPS*N
Heterosk	NP-C/R	**0.20**	**-230**	0.23	**-272**	**0.20**	**-228**
Homog		10%**	-81***	-2%	-41***	8%*	-74***
Homosk	NP-C	7%**	-66***	2%	-17**	9%	-52***
Heterosk	Flat	22%***	-42***	44%***	-701***	102%***	-309***
	Param	4%*	-60***	35%***	-135***	7%	-52***
	NP-disc	1%	-9**	**-7%**	-1	2%	-20***
	NP-R	1%	-5*	28%***	-63***	3%	-16***
	NP-C	3%*	-6*	3%	-5**	0.1%	-5**

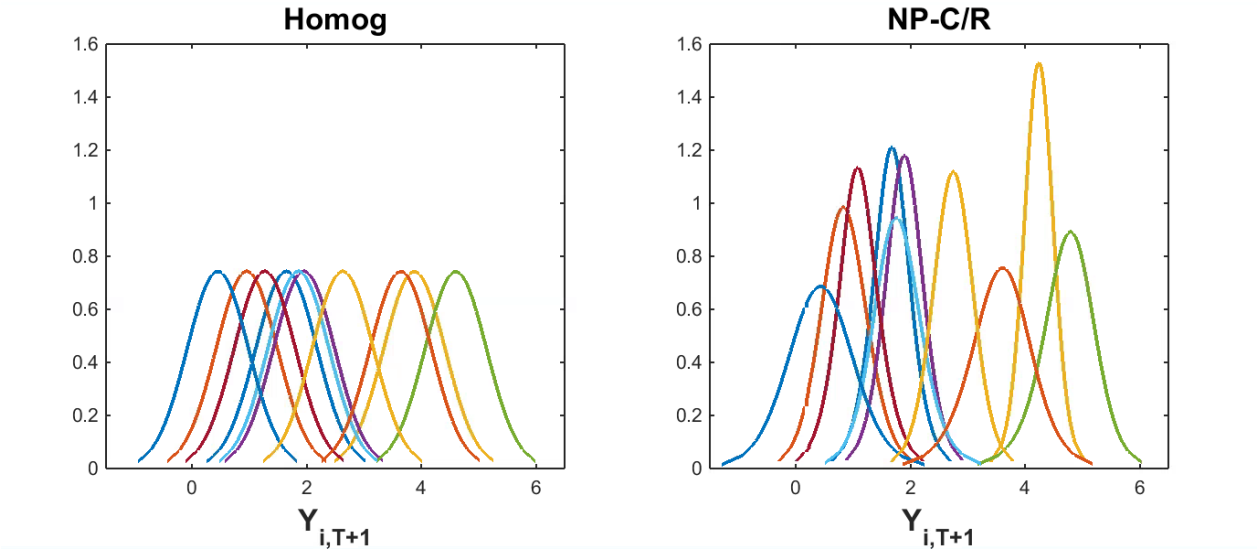
The point forecasts are evaluated by the Mean Square Error (MSE) together with the Diebold and Mariano (1995) test. The performance of the density forecasts is assessed by the log predictive score (LPS) and the Amisano and Giacomini (2007) test. For the benchmark predictor Heterosk-NP-C/R, the table reports the exact values of MSE and LPS. For other predictors, the table reports the percentage deviations from the benchmark MSE and difference with respect to the benchmark LPS. The tests are conducted with respect to the benchmark, with significance levels indicated by *: 10%, **: 5%, ***: 1%. The entries in bold indicate the best predictor in each column.

Figure 7.2: PIT



Red lines indicate the confidence interval.

Figure 7.3: Predictive Distributions: 10 Randomly Selected Firms

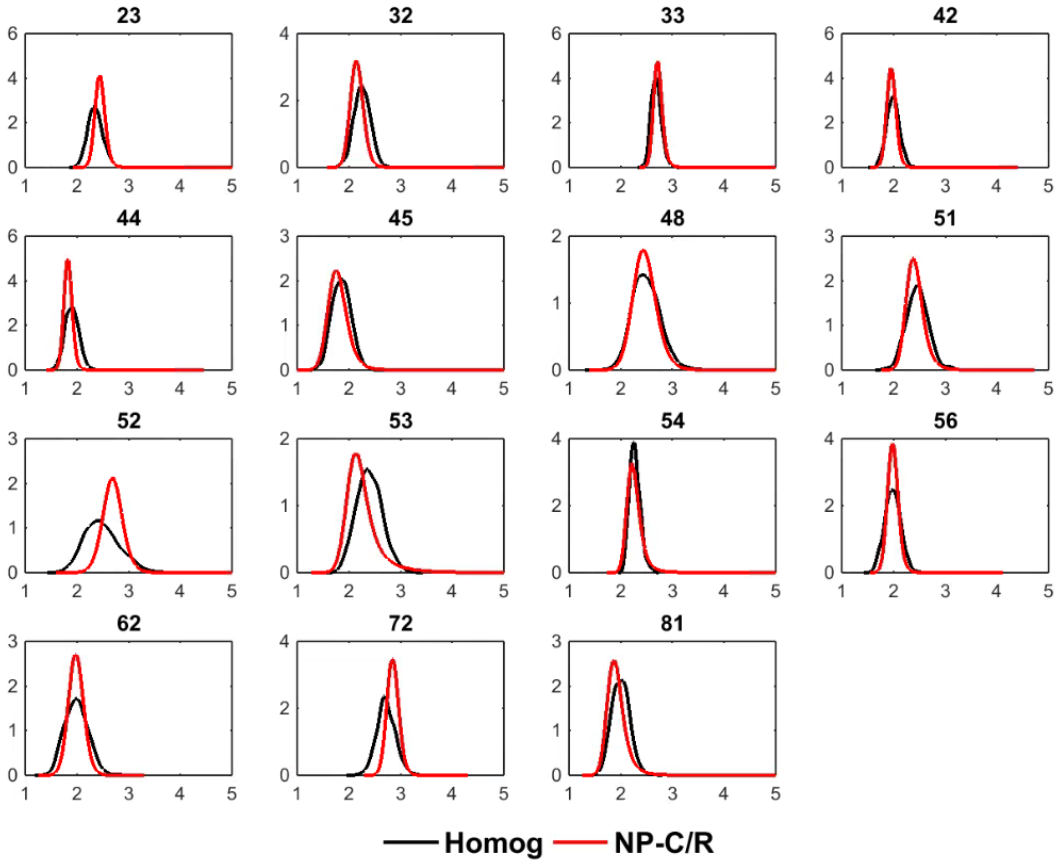


terms of the "Homog" predictor, all predictive distributions share the same Gaussian shape paralleling with each other. On the contrary, in terms of the "NP-C/R" predictor, it is clear that the predictive distributions are fairly different in the center location, variance, and skewness.

Figure 7.4 further aggregates the predictive distributions over sectors based on two-digit NAICS codes (Table 7.4). It plots the predictive distributions of the log of the average employment within each sector. Comparing "Homog" and "NP-C/R" across sectors, we can see the following several patterns. First, "NP-C/R" predictive distributions tend to be narrower. The reason is that "NP-C/R" tailors to each individual firm while "Homog" prescribes a general model to all the firms, so "NP-C/R" yields more precise predictive distributions. Second, "NP-C/R" predictive distributions have longer right tails, whereas "Homog" ones are distributed in the standard bell shape. The long right tails in "NP-C/R" concur with the general intuition that good ideas are scarce. Finally, there are substantial heterogeneities in density forecasts across sectors. For sectors with relatively large average employment, e.g. "construction" (sector 23), "Homog" pushes the forecasts down, hence systematically underpredicts their future employment, while "NP-C/R" respects this source of heterogeneity and significantly lessens the underprediction problem. On the other hand, for sectors with relatively small average employment, e.g. "Retail Trade" (sector 44), "Homog" introduces an upward bias into the forecasts, while "NP-C/R" reduces such bias by flexibly estimating the underlying distribution of firm-specific heterogeneities.

The latent heterogeneity structure is presented in Figure 7.5, which plots the joint distributions of the estimated individual effects and the conditional variable in the R&D setup. In all the three subgraphs, the pairwise relationships among $\lambda_{i,\text{level}}$, $\lambda_{i,\text{RD}}$, and standardized y_{i0} are nonlinear and exhibit multiple components, which reassures the utilization of nonparametric prior with correlated

Figure 7.4: Predictive Distributions: Aggregated by Sectors

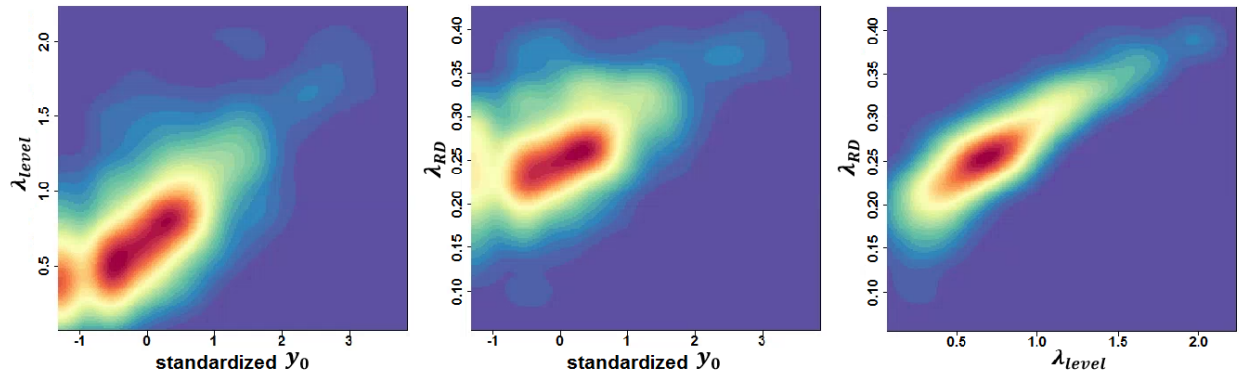


Subgraph titles are two-digit NAICS codes. Only sectors with more than 10 firms are shown.

Table 7.4: Two-digit NAICS Codes

Code	Sector
11	Agriculture, Forestry, Fishing and Hunting
21	Mining, Quarrying, and Oil and Gas Extraction
22	Utilities
23	Construction
31-33	Manufacturing
42	Wholesale Trade
44-45	Retail Trade
48-49	Transportation and Warehousing
51	Information
52	Finance and Insurance
53	Real Estate and Rental and Leasing
54	Professional, Scientific, and Technical Services
56	Administrative and Support and Waste Management and Remediation Services
61	Educational Services
62	Health Care and Social Assistance
71	Arts, Entertainment, and Recreation
72	Accommodation and Food Services
81	Other Services (except Public Administration)

Figure 7.5: Joint Distributions of $\hat{\lambda}_i$ and Condition Variable



random coefficients. Furthermore, $\lambda_{i,\text{level}}$, $\lambda_{i,\text{RD}}$, and standardized y_{i0} are positively correlated with each other, which roughly indicates that larger firms respond more positively to R&D activities within the KFS young firm sample.[28]

[28]The model here mainly serves the forecasting purpose, so we need to be careful with any causal interpretation.

8 Concluding Remarks

This paper proposes a semiparametric Bayesian predictor which performs well in density forecasts of individuals in a panel data setup. It considers the underlying distribution of individual effects and pools the information from the whole cross-section in an efficient and flexible way. Monte Carlo simulations and an empirical application to young firm dynamics show that the keys for the better density forecasts are, in order of importance, nonparametric Bayesian prior, cross-sectional heteroskedasticity, and correlated random coefficients.

Moving forward, I plan to extend my research in the following several directions:

Theoretically, I will continue the Bayesian asymptotic discussion with strong posterior consistency and rates of convergence.

Methodologically, I will explore some variations of the current setup. First, some empirical studies may include a large number of covariates with potential heterogeneous effects (i.e. more variables included in $w_{i,t-1}$), so it is both theoretically and empirically desirable to investigate a variable selection scheme in a high-dimensional nonparametric Bayesian framework. Chung and Dunson (2012) and Liverani *et al.* (2015) employ variable selection via binary switches, which may be adaptable to the panel data setting. Another possible direction is to construct a Bayesian-Lasso-type estimator coherent with the current nonparametric Bayesian implementation. Second, I will consider panel VAR (Canova and Ciccarelli, 2013), a useful tool to incorporate several variables for each of the individuals and to jointly model the evolution of these variables, allowing us to take more information into account for forecasting purposes and offer richer insights into the latent heterogeneity structure. Meanwhile, it is also interesting to incorporate extra cross-variable restrictions derived from economic theories and implement the Bayesian GMM method as proposed in Shin (2014). Third, I will experiment with nonlinear panel data models, such as the Tobit model that helps accommodate firms' endogenous exit choice. Such extension would be numerically feasible, but requires further theoretical work. A natural next step would be extending the theoretical discussion to the family of "generalized linear models".

References

Akcigit, U. and Kerr, W. R. (2010). Growth through heterogeneous innovations.

Amewou-Atisso, M., Ghosal, S., Ghosh, J. K. and Ramamoorthi, R. V. (2003). Posterior consistency for semi-parametric regression problems. *Bernoulli*, **9** (2), 291–312.

Amisano, G. and Geweke, J. (2016). Prediction using several macroeconomic models. *The Review of Economics and Statistics*, forthcoming.

— and Giacomini, R. (2007). Comparing density forecasts via weighted likelihood ratio tests. *Journal of Business & Economic Statistics*, **25** (2), 177–190.

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, pp. 1152–1174.

Arellano, M. and Bonhomme, S. (2012). Identifying distributional characteristics in random coefficients panel data models. *The Review of Economic Studies*, **79** (3), 987–1020.

— and Bover, O. (1995). Another look at the instrumental variable estimation of error-components models. *Journal of Econometrics*, **68** (1), 29 – 51.

Atchadé, Y. F. and Rosenthal, J. S. (2005). On adaptive Markov chain Monte Carlo algorithms. *Bernoulli*, **11** (5), 815–828.

Basu, S. and Chib, S. (2003). Marginal likelihood and Bayes factors for Dirichlet process mixture models. *Journal of the American Statistical Association*, **98** (461), 224–235.

Blackwell, D. and Dubins, L. (1962). Merging of opinions with increasing information. *The Annals of Mathematical Statistics*, **33** (3), 882–886.

Burda, M. and Harding, M. (2013). Panel probit with flexible correlated effects: quantifying technology spillovers in the presence of latent heterogeneity. *Journal of Applied Econometrics*, **28** (6), 956–981.

—, — and Hausman, J. (2012). A Poisson mixture model of discrete choice. *Journal of Econometrics*, **166** (2), 184–203.

Canova, F. and Ciccarelli, M. (2013). *Panel Vector Autoregressive Models: A Survey*. Working Paper Series, European Central Bank 1507, European Central Bank.

Chung, Y. and Dunson, D. B. (2012). Nonparametric bayes conditional distribution modeling with variable selection. *Journal of the American Statistical Association*.

Delaigle, A., Hall, P. and Meister, A. (2008). On deconvolution with repeated measurements. *The Annals of Statistics*, pp. 665–685.

Diaconis, P. and Freedman, D. (1986). On inconsistent Bayes estimates of location. *The Annals of Statistics*, pp. 68–87.

Diebold, F. X., Gunther, T. A. and Tay, A. S. (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review*, **39** (4), 863–883.

— and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, **13** (3).

Doob, J. L. (1949). Application of the theory of martingales. *Le calcul des probabilites et ses applications*, pp. 23–27.

Dunson, D. B. (2009). Nonparametric Bayes local partition models for random effects. *Biometrika*, **96** (2), 249–262.

— and Park, J.-H. (2008). Kernel stick-breaking processes. *Biometrika*, **95** (2), 307–323.

Efron, B. (2012). *Large-scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*, vol. 1. Cambridge University Press.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, **90** (430), 577–588.

Fernandez, C. and Steel, M. F. (2000). Bayesian regression analysis with scale mixtures of normals. *Econometric Theory*, **16** (01), 80–101.

Freedman, D. A. (1963). On the asymptotic behavior of Bayes' estimates in the discrete case. *The Annals of Mathematical Statistics*, pp. 1386–1403.

— (1965). On the asymptotic behavior of Bayes estimates in the discrete case II. *The Annals of Mathematical Statistics*, **36** (2), 454–456.

Galambos, J. and Simonelli, I. (2004). *Products of Random Variables: Applications to Problems of Physics and to Arithmetical Functions*. Marcel Dekker.

Geweke, J. and Amisano, G. (2010). Comparing and evaluating Bayesian predictive distributions of asset returns. *International Journal of Forecasting*, **26** (2), 216–230.

Ghosal, S., Ghosh, J. K., Ramamoorthi, R. *et al.* (1999). Posterior consistency of Dirichlet mixtures in density estimation. *The Annals of Statistics*, **27** (1), 143–158.

Ghosh, J. K. and Ramamoorthi, R. (2003). *Bayesian Nonparametrics*. Springer-Verlag.

Griffin, J. E. (2016). An adaptive truncation method for inference in Bayesian nonparametric models. *Statistics and Computing*, **26** (1), 423–441.

Gu, J. and Koenker, R. (2015). Unobserved heterogeneity in income dynamics: an empirical Bayes perspective. *Journal of Business & Economic Statistics*, forthcoming.

— and — (2016). Empirical Bayesball remixed: empirical Bayes methods for longitudinal data. *Journal of Applied Econometrics*, pp. n/a–n/a.

Hall, B. H. and Rosenberg, N. (2010). *Handbook of the Economics of Innovation*, vol. 1. Elsevier.

Haltiwanger, J., Jarmin, R. S. and Miranda, J. (2012). Who creates jobs? Small versus large versus young. *Review of Economics and Statistics*, **95** (2), 347–361.

Hastie, D. I., Liverani, S. and Richardson, S. (2015). Sampling from Dirichlet process mixture models with unknown concentration parameter: mixing issues in large data implementations. *Statistics and Computing*, **25** (5), 1023–1037.

Hirano, K. (2002). Semiparametric Bayesian inference in autoregressive panel data models. *Econometrica*, **70** (2), 781–799.

Hjort, N. L., Holmes, C., Müller, P. and Walker, S. G. (2010). *Bayesian Nonparametrics*. Cambridge University Press.

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, **96** (453), 161–173.

— and — (2002). Approximate Dirichlet process computing in finite normal mixtures: smoothing and prior information. *Journal of Computational and Graphical Statistics*, **11** (3), 508–532.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, Berkeley, Calif.: University of California Press, pp. 361–379.

Jensen, M. J., Fisher, M. and Tkac, P. (2015). Mutual fund performance when learning the distribution of stock-picking skill.

Kalli, M., Griffin, J. E. and Walker, S. G. (2011). Slice sampling mixture models. *Statistics and Computing*, **21** (1), 93–105.

Kelker, D. (1970). Distribution theory of spherical distributions and a location-scale parameter generalization. *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 419–430.

Lee, Y., Amaral, L. A. N., Canning, D., Meyer, M. and Stanley, H. E. (1998). Universal features in the growth dynamics of complex organizations. *Physical Review Letters*, **81** (15), 3275.

Liu, L., Moon, H. R. and Schorfheide, F. (2016). Forecasting with dynamic panel data models.

Liverani, S., Hastie, D. I., Azizi, L., Papathomas, M. and Richardson, S. (2015). PReMiuM: an R package for profile regression mixture models using Dirichlet processes. *Journal of Statistical Software*, **64** (7).

Marcellino, M., Stock, J. H. and Watson, M. W. (2006). A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series. *Journal of Econometrics*, **135** (1), 499–526.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, **9** (2), 249–265.

Norets, A. (2010). Approximation of conditional densities by smooth mixtures of regressions. *The Annals of Statistics*, **38** (3), 1733–1766.

— and Pelenis, J. (2012). Bayesian modeling of joint and conditional distributions. *Journal of Econometrics*, **168** (2), 332–346.

— and — (2014). Posterior consistency in conditional density estimation by covariate dependent mixtures. *Econometric Theory*, **30**, 606–646.

Papaspiliopoulos, O. and Roberts, G. O. (2008). Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika*, **95** (1), 169–186.

Pati, D., Dunson, D. B. and Tokdar, S. T. (2013). Posterior consistency in conditional distribution estimation. *Journal of Multivariate Analysis*, **116**, 456–472.

Pav, S. E. (2015). Moments of the log non-central chi-square distribution. *arXiv preprint arXiv:1503.06266*.

Pelenis, J. (2014). Bayesian regression with heteroscedastic error density and parametric mean function. *Journal of Econometrics*, **178**, 624–638.

Robb, A., Ballou, J., DesRoches, D., Potter, F., Zhao, Z. and Reedy, E. (2009). An overview of the Kauffman Firm Survey: results from the 2004-2007 data. *Available at SSRN 1392292*.

— and Seamans, R. (2014). The role of R&D in entrepreneurial finance and performance. *Available at SSRN 2341631*.

Robbins, H. (1956). An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, Berkeley and Los Angeles.

Rossi, P. E. (2014). *Bayesian Non- and Semi-parametric Methods and Applications*. Princeton University Press.

Santarelli, E., Klomp, L. and Thurik, A. R. (2006). Gibrat's law: an overview of the empirical literature. In *Entrepreneurship, Growth, and Innovation*, Springer, pp. 41–73.

Schwartz, L. (1965). On Bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, **4** (1), 10–26.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, pp. 639–650.

Shin, M. (2014). Bayesian GMM.

Tokdar, S. T. (2006). Posterior consistency of Dirichlet location-scale mixture of normals in density estimation and regression. *Sankhyā: The Indian Journal of Statistics*, pp. 90–110.

Walker, S. G. (2007). Sampling the Dirichlet mixture model with slices. *Communications in Statistics - Simulation and Computation*, **36** (1), 45–54.

Wu, Y. and Ghosal, S. (2008). Kullback Leibler property of kernel mixture priors in Bayesian density estimation. *Electronic Journal of Statistics*, **2**, 298–331.

Yau, C., Papaspiliopoulos, O., Roberts, G. O. and Holmes, C. (2011). Bayesian non-parametric hidden Markov models with applications in genomics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **73** (1), 37–57.

Zarutskie, R. and Yang, T. (2015). How did young firms fare during the great recession? Evidence from the Kauffman Firm Survey. In *Measuring Entrepreneurial Businesses: Current Knowledge and Challenges*, University of Chicago Press.

A Notations

$U(a,b)$ represents a **uniform distribution** with minimum value a and maximum value b. If $a=0$ and $b=1$, we obtain the standard uniform distribution, $U(0,1)$.

$N\left(\mu,\sigma^2\right)$ or $N\left(x;\mu,\sigma^2\right)$ stands for a **Gaussian distribution** with mean μ and variance σ^2. Its probability distribution function (pdf) is given by $\phi\left(x;\mu,\sigma^2\right)$. When $\mu=0$ and $\sigma^2=1$ (i.e. standard normal), we reduce the notation to $\phi(x)$. The corresponding cumulative distribution functions (cdf) are denoted as $\Phi\left(x;\mu,\sigma^2\right)$ and $\Phi(x)$, respectively. The same convention holds for multivariate normal, where $N(\mu,\Sigma)$, $N(x;\mu,\Sigma)$, $\phi(x;\mu,\Sigma)$, and $\Phi(x;\mu,\Sigma)$ are for the distribution with the mean vector μ and the covariance matrix Σ.

$TN\left(\mu,\sigma^2,a,b\right)$ denotes a **truncated normal distribution** with μ and σ^2 being the mean and variance before truncation, and a and b being the lower and upper end of the truncated interval.

The **gamma distribution** is denoted as $\mathrm{Ga}(x;a,b)$ with pdf being $f_{\mathrm{Ga}}(x;a,b)=\frac{b^a}{\Gamma(a)}x^{a-1}e^{-bx}$. The according **inverse-gamma distribution** is given by $\mathrm{IG}(x;a,b)$ with pdf being $f_{\mathrm{IG}}(x;a,b)=\frac{b^a}{\Gamma(a)}x^{-a-1}e^{-b/x}$. The $\Gamma(\cdot)$ in the denominators is the gamma function.

The **inverse Wishart distribution** is a generalization of the inverse gamma distribution to multi-dimensional setups. Let Ω be a $d\times d$ matrix, then the inverse Wishart distribution is denoted as $\mathrm{IW}(\Omega;\Psi,\nu)$, and its pdf is $f_{\mathrm{IW}}(\Omega;\Psi,\nu)=\frac{|\Psi|^{\frac{\nu}{2}}}{2^{\frac{\nu d}{2}}\Gamma_d\left(\frac{\nu}{2}\right)}|\Omega|^{-\frac{\nu+d+1}{2}}e^{-\frac{1}{2}tr(\Psi\Omega^{-1})}$. When Ω is a scalar, the inverse Wishart distribution is reduced to a inverse-gamma distribution with $a=\nu/2,\ b=\Psi/2$.

$\mathbf{1}(\cdot)$ is an **indicator function** that equals 1 if the condition in the parenthesis is satisfied and equals 0 otherwise.

I_N is an $N\times N$ **identity matrix**.

In the **panel data** setup, for a generic variable z, which can be $v,\ w\ ,x,$ or y, z_{it} is a $d_z\times 1$ vector, and $z_{i,t_1:t_2}=(z_{it_1},\cdots,z_{it_2})$ is a $d_z\times(t_2-t_1+1)$ matrix.

$\|\cdot\|$ represents the **Euclidean norm**, i.e. for a n-dimensional vector $z=[z_1,z_2,\cdots,z_n]'$, $\|z\|=\sqrt{z_1^2+\cdots+z_n^2}$.

$\mathrm{supp}(\cdot)$ denotes the **support** of a probability measure.

B Algorithms

B.1 Hyperparameters

Recall the prior for the common parameters:

$$\left(\beta,\sigma^2\right)\sim N\left(m_0^\beta,\psi_0^\beta\sigma^2\right)\mathrm{IG}\left(\sigma^2;\ a_0^{\sigma^2},b_0^{\sigma^2}\right).$$

The hyperparameters are chosen in a relatively ignorant sense without inferring too much from the data except aligning the scale according to the variance of the data.

$$a_0^{\sigma^2} = 2, \tag{B.1}$$

$$b_0^{\sigma^2} = \hat{E}^i\left(\widehat{Var}_i^t\left(y_{it}\right)\right)\cdot\left(a_0^{\sigma^2}-1\right) = \hat{E}^i\left(\widehat{Var}_i^t\left(y_{it}\right)\right), \tag{B.2}$$

$$m_0^{\beta} = 0.5, \tag{B.3}$$

$$\psi_0^{\beta} = \frac{1}{b_0^{\sigma^2}/\left(a_0^{\sigma^2}-1\right)} = \frac{1}{\hat{E}^i\left(\widehat{Var}_i^t\left(y_{it}\right)\right)}. \tag{B.4}$$

In equation (B.2) here and equation (B.5) below, $\hat{E}_i^t$ and $\widehat{Var}_i^t$ stand for the sample mean and variance for firm i over $t = 1, \cdots, T$, and $\hat{E}^i$ and $\widehat{Var}^i$ are the sample mean and variance over the whole cross-section $i = 1, \cdots, N$. Equation (B.2) ensures that on average the prior and the data have a similar scale. Equation (B.3) conjectures that the young firm dynamics are highly likely persistent and stationary. Since we don't have strong prior information in the common parameters, their priors are chosen to be not very restrictive. Equation (B.1) characterizes a rather less informative prior on σ^2 with infinite variance, and Equation (B.4) assumes that the prior variance of β is equal to 1 on average.

The hyperpriors for the DPM prior are specified as:

$$G_0\left(\mu_k, \omega_k^2\right) = N\left(\mu_k;\ m_0^{\lambda}, \psi_0^{\lambda}\omega_k^2\right)\mathrm{IG}\left(\omega_k^2;\ a_0^{\lambda}, b_0^{\lambda}\right),$$
$$\alpha \sim \mathrm{Ga}\left(\alpha;\ a_0^{\alpha}, b_0^{\alpha}\right).$$

Similarly, the hyperparameters are chosen to be:

$$a_0^{\lambda} = 2,\ b_0^{\lambda} = \widehat{Var}^i\left(\hat{E}_i^t\left(y_{it}\right)\right)\cdot\left(a_0^{\lambda}-1\right) = \widehat{Var}^i\left(\hat{E}_i^t\left(y_{it}\right)\right), \tag{B.5}$$

$$m_0^{\lambda} = 0,\ \psi_0^{\lambda} = 1,$$

$$a_0^{\alpha} = 2,\ b_0^{\alpha} = 2. \tag{B.6}$$

where b_0^{λ} is selected to match the scale, while a_0^{λ}, m_0^{λ}, and ψ_0^{λ} yields a relatively ignorant and diffuse prior. Following Ishwaran and James (2001, 2002), the hyperparameters for the DP scale parameter α in equation (B.6) allows for a flexible component structure with a wide range of component numbers. The truncated number of components is set to be $K = 50$, so that the approximation error is uniformly bounded by Ishwaran and James (2001) Theorem 2:

$$\left\| f^{\lambda,K} - f^{\lambda} \right\| \sim 4N\exp\left(-\frac{K-1}{\alpha}\right) \leq 2.10\times 10^{-18},$$

at the prior mean of α ($\bar{\alpha} = 1$) and cross-sectional sample size $N = 1000$.

I have also examined other choices of hyperparameters, and results are not very sensitive to hyperparameters as long as the implied priors are flexible enough to cover the range of observables.

B.2 Random-Walk Metropolis-Hastings

When there is no closed-form conditional posterior distribution in some MCMC steps, it is helpful to employ the Metropolis-within-Gibbs sampler and use the random-walk Metropolis-Hastings (RWMH) algorithm for those steps. The adaptive RWMH algorithm below is based on Atchadé and Rosenthal (2005) and Griffin (2016), which adaptively adjust the random walk step size in order to keep acceptance rates around certain desirable percentage.

Algorithm B.1. *(Adaptive RWMH)*

Let us consider a generic variable θ. For each iteration $s = 1, \cdots, n_{sim}$,

1. *Draw candidate $\tilde{\theta}$ from the random-walk proposal density $\tilde{\theta} \sim N\left(\theta^{(s-1)}, \zeta^{(s)}\Sigma\right)$.*
2. *Calculate the acceptance rate*

$$a.r.(\tilde{\theta}|\theta^{(s-1)}) = \min\left(1, \frac{p(\tilde{\theta}|\cdot)}{p(\theta^{(s-1)}|\cdot)}\right),$$

   where $p(\theta|\cdot)$ is the conditional posterior distribution of interest.
3. *Accept the proposal and set $\theta^{(s)} = \tilde{\theta}$ with probability $a.r.(\tilde{\theta}|\theta^{(s-1)})$. Otherwise, reject the proposal and set $\theta^{(s)} = \theta^{(s-1)}$.*
4. *Update the random-walk step size for the next iteration,*

$$\log \zeta^{(s+1)} = \rho\left(\log \zeta^{(s)} + s^{-c}\left(a.r.(\tilde{\theta}|\theta^{(s-1)}) - a.r.^{\star}\right)\right),$$

   where $0.5 < c \leq 1$, $a.r.^{\star}$ is the target acceptance rate, and

$$\rho(x) = \min(|x|, \bar{x}) \cdot sgn(x),$$

   where $\bar{x} > 0$ is a very large number.

Remark B.2. (i) In step 1, since the algorithms in this paper only consider RWMH on conditionally independent scalar variables, Σ is simply taken to be 1.

(ii) In step 4, I choose $c = 0.55$, $\text{a.r.}^{\star} = 30\%$ in the numerical exercises, following Griffin (2016).

B.3 Details on Posterior Samplers

The formulas below focus on the (correlated) random coefficients model in Algorithms 5.1 and 5.2 where the (correlated) random effects model in Algorithms 3.1 and 3.2 are special cases with solely univariate λ_i.

B.3.1 Step 2: Component Parameters

Random Coefficients Model For $z = \lambda,\ l$ and $k^z = 1, \cdots, K^z$, draw $\left(\mu_{k^z}^{z(s)}, \Omega_{k^z}^{z(s)}\right)$ from a multivariate-normal-inverse-Wishart distribution (or a normal-inverse-gamma distribution if z is a scalar) $p\left(\mu_{k^z}^{z(s)}, \Omega_{k^z}^{z(s)} \middle| \left\{z_i^{(s-1)}\right\}_{i \in J_{k^z}^{z(s-1)}}\right)$:

$$
\begin{aligned}
\left(\mu_{k^z}^{z(s)}, \Omega_{k^z}^{z(s)}\right) &\sim N\left(\mu_{k^z}^{z(s)};\ m_{k^z}^z, \psi_{k^z}^z \Omega_{k^z}^{z(s)}\right) \mathrm{IW}\left(\Omega_{k^z}^{z(s)};\ \Psi_{k^z}^z, \nu_{k^z}^z\right), \\
\hat{m}_{k^z}^z &= \frac{1}{n_{k^z}^{z(s-1)}} \sum_{i \in J_{k^z}^{z(s-1)}} z_i^{(s-1)}, \\
\psi_{k^z}^z &= \left((\psi_0^z)^{-1} + n_{k^z}^{z(s-1)}\right)^{-1}, \\
m_{k^z}^z &= \psi_{k^z}^z \left((\psi_0^z)^{-1} m_0^z + \sum_{i \in J_{k^z}^{z(s-1)}} z_i^{(s-1)}\right), \\
\nu_{k^z}^z &= \nu_0^z + n_{k^z}^{z(s-1)}, \\
\Psi_{k^z}^z &= \Psi_0^z + \sum_{i \in J_{k^z}^{z(s-1)}} \left(z_i^{(s-1)}\right)^2 + m_0^{z\prime} (\psi_0^z)^{-1} m_0^z - m_{k^z}^{z\prime} (\psi_{k^z}^z)^{-1} m_{k^z}^z.
\end{aligned}
$$

Correlated Random Coefficients Model Due to the complexity arising from the conditional structure, I break the updating procedure for $\left(\mu_{k^z}^{z(s)}, \Omega_{k^z}^{z(s)}\right)$ into two steps. For $z = \lambda,\ l$ and $k^z = 1, \cdots, K^z$,

(a) Draw $\mu_{k^z}^{z(s)}$ from a matricvariate-normal distribution (or a multivariate-normal distribution if z is a scalar) $p\left(\mu_{k^z}^{z(s)} \middle| \Omega_{k^z}^{z(s-1)}, \left\{z_i^{(s-1)}, c_{i0}\right\}_{i \in J_{k^z}^{z(s-1)}}\right)$:

$$
\begin{aligned}
\mathrm{vec}\left(\mu_{k^z}^{z(s)}\right) &\sim N\left(\mathrm{vec}\left(\mu_{k^z}^{z(s)}\right);\ \mathrm{vec}\left(m_{k^z}^z\right), \psi_{k^z}^z\right), \\
\hat{m}_{k^z}^{z,zc} &= \sum_{i \in J_{k^z}^{z(s-1)}} z_i^{(s-1)} \left[1, c_{i0}'\right], \\
\hat{m}_{k^z}^{z,cc} &= \sum_{i \in J_{k^z}^{z(s-1)}} \left[1, c_{i0}'\right]' \left[1, c_{i0}'\right], \\
\hat{m}_{k^z}^z &= \hat{m}_{k^z}^{z,zc} \left(\hat{m}_{k^z}^{z,cc}\right)^{-1}, \\
\psi_{k^z}^z &= \left[(\psi_0^z)^{-1} + \hat{m}_{k^z}^{z,cc} \otimes \left(\Omega_{k^z}^{z(s-1)}\right)^{-1}\right]^{-1}, \\
\mathrm{vec}\left(m_{k^z}^z\right) &= \psi_{k^z}^z \left[(\psi_0^z)^{-1} \mathrm{vec}\left(m_0^z\right) + \left(\hat{m}_{k^z}^{z,cc} \otimes \left(\Omega_{k^z}^{z(s-1)}\right)^{-1}\right) \mathrm{vec}\left(\hat{m}_{k^z}^z\right)\right],
\end{aligned}
$$

where $\mathrm{vec}(\cdot)$ denotes matrix vectorization, and $\otimes$ is the Kronecker product.

(b) Draw $\Omega_{k^z}^{z(s)}$ from an inverse-Wishart distribution (or an inverse-gamma distribution if z is a

scalar) $p\left(\Omega_{k^z}^{z(s)} \middle| \mu_{k^z}^{z(s)}, \left\{z_i^{(s-1)}, c_{i0}\right\}_{i \in J_{k^z}^{z(s-1)}}\right)$:

$$\begin{aligned}
\Omega_{k^z}^{z(s)} &\sim \mathrm{IW}\left(\Omega_{k^z}^{z(s)};\ \Psi_{k^z}^{z}, \nu_{k^z}^{z}\right), \\
\nu_{k^z}^{z} &= \nu_0^z + n_{k^z}^{z(s-1)}, \\
\Psi_{k^z}^{z} &= \Psi_0^z + \sum_{i \in J_{k^z}^{z(s-1)}} \left(z_i^{(s-1)} - \mu_{k^z}^{z(s)}\left[1, c_{i0}'\right]'\right)\left(z_i^{(s-1)} - \mu_{k^z}^{z(s)}\left[1, c_{i0}'\right]'\right)'.
\end{aligned}$$

B.3.2 Step 4: Individual-specific Parameters

For $i = 1, \cdots, N$, draw $\lambda_i^{(s)}$ from a multivariate-normal distribution (or a normal distribution if λ is a scalar) $p\left(\lambda_i^{(s)} \middle| \mu_{\gamma_i^\lambda}^{\lambda(s)}, \Omega_{\gamma_i^\lambda}^{\lambda(s)}, \left(\sigma_i^2\right)^{(s-1)}, \beta^{(s-1)}, D_i, D_A\right)$:

$$\begin{aligned}
\lambda_i^{(s)} &\sim N\left(m_i^\lambda, \Sigma_i^\lambda\right), \\
\Sigma_i^\lambda &= \left(\left(\Omega_{\gamma_i^\lambda}^{\lambda(s)}\right)^{-1} + \left(\left(\sigma_i^2\right)^{(s-1)}\right)^{-1} \sum_{t=t_{0i}}^{t_{1i}} w_{i,t-1} w_{i,t-1}'\right)^{-1}, \\
m_i^\lambda &= \Sigma_i^\lambda \left(\left(\Omega_{\gamma_i^\lambda}^{\lambda(s)}\right)^{-1} \tilde{\mu}_i^\lambda + \left(\left(\sigma_i^2\right)^{(s-1)}\right)^{-1} \sum_{t=t_{0i}}^{t_{1i}} w_{i,t-1}\left(y_{it} - \beta^{(s-1)\prime} x_{i,t-1}\right)\right),
\end{aligned}$$

where the conditional "prior" mean is characterized by

$$\tilde{\mu}_i^\lambda = \begin{cases} \mu_{\gamma_i^\lambda}^{\lambda(s)}, & \text{for the random coefficients model,} \\ \mu_{\gamma_i^\lambda}^{\lambda(s)}\left[1, c_{i0}'\right]', & \text{for the correlated random coefficients model.} \end{cases}$$

B.3.3 Step 5: Common parameters

Cross-sectional Homoskedasticity Draw $\left(\beta^{(s)}, \sigma^{2(s)}\right)$ from a linear regression model with "unknown" variance, $p\left(\beta^{(s)}, \sigma^{2(s)} \middle| \left\{\lambda_i^{(s)}\right\}, D\right)$:

$$\begin{aligned}
\left(\beta^{(s)}, \sigma^{2(s)}\right) &\sim N\left(\beta^{(s)};\, m^{\beta}, \psi^{\beta}\sigma^{2(s)}\right) \mathrm{IG}\left(\sigma^{2(s)};\, a^{\sigma^2}, b^{\sigma^2}\right), \\
\psi^{\beta} &= \left(\left(\psi_0^{\beta}\right)^{-1} + \sum_{i=1}^{N}\sum_{t=t_{0i}}^{t_{1i}} x_{i,t-1}x_{i,t-1}'\right)^{-1}, \\
m^{\beta} &= \psi^{\beta}\left(\left(\psi_0^{\beta}\right)^{-1} m_0^{\beta} + \sum_{i=1}^{N}\sum_{t=t_{0i}}^{t_{1i}} x_{i,t-1}\left(y_{it} - \lambda_i^{(s)\prime} w_{i,t-1}\right)\right), \\
a^{\sigma^2} &= a_0^{\sigma^2} + \frac{NT}{2} \\
b^{\sigma^2} &= b_0^{\sigma^2} + \frac{1}{2}\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it} - \lambda_i^{(s)\prime} w_{i,t-1}\right)^2 + m_0^{\beta\prime}\left(\psi_0^{\beta}\right)^{-1} m_0^{\beta} - m^{\beta\prime}\left(\psi^{\beta}\right)^{-1} m^{\beta}\right).
\end{aligned}$$

Cross-sectional Heteroskedasticity Draw $\beta^{(s)}$ from a linear regression model with "known" variance, $p\left(\beta^{(s)} \middle| \left\{\lambda_i^{(s)}, \left(\sigma_i^2\right)^{(s)}\right\}, D\right)$:

$$\begin{aligned}
\beta^{(s)} &\sim N\left(m^{\beta}, \Sigma^{\beta}\right), \\
\Sigma^{\beta} &= \left(\left(\Sigma_0^{\beta}\right)^{-1} + \left(\left(\sigma_i^2\right)^{(s)}\right)^{-1}\sum_{i=1}^{N}\sum_{t=t_{0i}}^{t_{1i}} x_{i,t-1}x_{i,t-1}'\right)^{-1}, \\
m^{\beta} &= \Sigma^{\beta}\left(\left(\Sigma_0^{\beta}\right)^{-1} m_0^{\beta} + \left(\left(\sigma_i^2\right)^{(s)}\right)^{-1}\sum_{i=1}^{N}\sum_{t=t_{0i}}^{t_{1i}} x_{i,t-1}\left(y_{it} - \lambda_i^{(s)\prime} w_{i,t-1}\right)\right).
\end{aligned}$$

Remark B.3. For unbalanced panels, the summations and products in steps 4 and 5 (Subsections B.3.2 and B.3.3) are instead over $t = t_{0i}, \cdots, t_{1i}$, the observed periods for individual i.

B.4 Slice-Retrospective Samplers

The next algorithm borrows the idea from some recent development in DPM sampling strategies (Dunson, 2009; Yau *et al.*, 2011; Hastie *et al.*, 2015), which integrates the slice sampler (Walker, 2007; Kalli *et al.*, 2011) and the retrospective sampler (Papaspiliopoulos and Roberts, 2008). By adding extra auxiliary variables, the sampler is able to avoid hard truncation in Ishwaran and James (2001, 2002). I experiment with it to check whether the approximation error due to truncation would significantly affect the density forecasts or not, and the results do not change much. The following algorithm is designed for the random coefficient case. A corresponding version for the correlated random coefficient case can be constructed in a similar manner.

The auxiliary variables u_i^z, $i = 1, \cdots, N$, are i.i.d. standard uniform random variables, i.e. $u_i^z \sim U(0,1)$. Then, the mixture of components in equation (2.6) can be rewritten as

$$z \sim \sum_{k^z=1}^{\infty} \mathbf{1}\left(u_i^z < p_{ik^z}^z\right) f^z\left(z;\ \theta_{k^z}^z\right),$$

where $z = \lambda,\ l$. By marginalizing over u_i^z, we can recover equation (2.6). Accordingly, we can define the number of active components as

$$K^{z,A} = \max_{1 \le i \le N} \gamma_i^z,$$

and the number of potential components (including active components) as

$$K^{z,P} = \min\left\{k^z :\ \left(1 - \sum_{j=1}^{k^z} p_j^z\right) < \min_{1 \le i \le N} u_i^z\right\}.$$

Although the number of components is infinite literally, we only need to care about the components that can potentially be occupied. Therefore, $K^{z,P}$ serves as an upper limit on the number of components that need to be updated at certain iteration. Here I suppress the iteration indicator s for exposition simplicity, but note that both $K^{z,A}$ and $K^{z,P}$ can change over iterations; this is indeed the highlight of this sampler.

Algorithm B.4. *(General Model: Random Coefficients III (Slice-Retrospective))*

For each iteration $s = 1, \cdots, n_{sim}$, steps 1-3 in Algorithm 5.1 are modified as follows:

For $z = \lambda,\ l$,

1. *Active components:*
   (a) *Number of active components:*

$$K^{z,A} = \max_{1 \le i \le N} \gamma_i^{z(s-1)}.$$

   (b) *Component probabilities: for $k^z = 1, \cdots, K^{z,A}$, draw $p_{k^z}^{z*}$ from the stick breaking process $p\left(\{p_{k^z}^{z*}\} \middle| \alpha^{z(s-1)}, \left\{n_{k^z}^{z(s-1)}\right\}\right)$:*

$$p_{k^z}^{z*} \sim SB\left(n_{k^z}^{z(s-1)},\ \alpha^{z(s-1)} + \sum_{j=k^z+1}^{K^{z,A}} n_j^{z(s-1)}\right), \quad k^z = 1, \cdots, K^{z,A}.$$

   (c) *Component parameters: for $k^z = 1, \cdots, K^{z,A}$, draw $\theta_{k^z}^{z*}$ from $p\left(\theta_{k^z}^{z*} \middle| \left\{z_i^{(s-1)}\right\}_{i \in J_{k^z}^{z(s-1)}}\right)$ as in Algorithm 3.1 step 2.*
   (d) *Label switching: jointly update $\left\{p_{k^z}^{z(s)}, \theta_{k^z}^{z(s)}, \gamma_i^{z*}\right\}_{k^z=1}^{K^{z,A}}$ based on $\left\{p_{k^z}^{z*}, \theta_{k^z}^{z*}, \gamma_i^{z(s-1)}\right\}_{k^z=1}^{K^{z,A}}$ by three Metropolis-Hastings label-switching moves:*

i. *randomly select two non-empty components, switch their component labels (γ_i^z), while leaving component parameters ($\theta_{k^z}^z$) and component probabilities ($p_{k^z}^z$) unchanged;*

ii. *randomly select two adjacent components, switch their component labels (γ_i^z) and component "stick lengths" ($\zeta_{k^z}^z$), while leaving component parameters ($\theta_{k^z}^z$) unchanged;*

iii. *randomly select two non-empty components, switch their component labels (γ_i^z) and component parameters ($\theta_{k^z}^z$), as well as update their component probabilities ($p_{k^z}^z$).*

Then, adjust $K^{z,A}$ accordingly.

2. *Auxiliary variables: for $i = 1, \cdots, N$, draw $u_i^{z(s)}$ from a uniform distribution $p\left(u_i^{z(s)} \left| \left\{p_{k^z}^{z(s)}\right\}, \gamma_i^{z*}\right.\right)$:*

$$u_i^{z(s)} \sim U\left(0,\ p_{\gamma_i^{z*}}^{z(s)}\right).$$

3. *DP scale parameter:*

   (a) *Draw the latent variable $\xi^{z(s)}$ from a beta distribution $p\left(\xi^{z(s)} \left| \alpha^{z(s-1)}, N\right.\right)$:*

$$\xi^{z(s)} \sim Beta\left(\alpha^{z(s-1)} + 1, N\right).$$

   (b) *Draw $\alpha^{z(s)}$ from a mixture of two gamma distributions $p\left(\alpha^{z(s)} \left| \xi^{z(s)}, K^{z,A}, N\right.\right)$:*

$$\begin{aligned}\alpha^{z(s)} \sim\ & p^{\alpha^z}\, Ga\left(\alpha^{z(s)};\ a^{\alpha^z} + K^{z,A},\ b^{\alpha^z} - \log \xi^{z(s)}\right) \\ & + \left(1 - p^{\alpha^z}\right) Ga\left(\alpha^{z(s)};\ a^{\alpha^z} + K^{z,A} - 1,\ b^{\alpha^z} - \log \xi^{z(s)}\right), \\ p^{\alpha^z} =\ & \frac{a^{\alpha^z} + K^{z,A} - 1}{N\left(b^{\alpha^z} - \log \xi^{z(s)}\right)}.\end{aligned}$$

4. *Potential components:*

   (a) *Component probabilities: start with $K^{z*} = K^{z,A}$,*

      i. *if $\left(1 - \sum_{j=1}^{K^{z*}} p_j^{z(s)}\right) < \min_{1 \le i \le N} u_i^{z(s)}$, set $K^{z,P} = K^{z*}$ and stop;*

      ii. *otherwise, let $K^{z*} = K^{z*}+1$, draw $\zeta_{K^{z*}}^z \sim Beta\left(1, \alpha^{z(s)}\right)$, update $p_{K^{z*}}^{z(s)} = \zeta_{K^{z*}}^z \prod_{j<K^{z*}} \left(1 - \zeta_j^z\right)$, and go to step (a-i).*

   (b) *Component parameters: for $k^z = K^{z,A} + 1, \cdots, K^{z,P}$, draw $\theta_{k^z}^{z(s)}$ from the DP base distribution G_0^z.*

5. *Component memberships: For $i = 1, \cdots N$, draw $\gamma_i^{z(s)}$ from a multinomial distribution $p\left(\left\{\gamma_i^{z(s)}\right\} \left| \left\{p_{k^z}^{z(s)}, \mu_{k^z}^{z(s)}, \Omega_{k^z}^{z(s)}\right\},\ u_i^{z(s)}, z_i^{(s-1)}\right.\right)$:*

$$\begin{aligned}&\gamma_i^{z(s)} = k^z,\ \textit{with probability}\ p_{ik^z}^z,\ k^z = 1, \cdots, K^{z,P}, \\ &p_{ik^z}^z \propto p_{k^z}^{z(s)} \phi\left(z_i^{(s-1)}; \mu_{k^z}^{z(s)}, \Omega_{k^z}^{z(s)}\right) \mathbf{1}\left(u_i^{z(s)} < p_{k^z}^{z(s)}\right), \quad \sum_{k^z=1}^{K^{z,P}} p_{ik^z}^z = 1.\end{aligned}$$

The remaining part of the algorithm resembles steps 4 and 5 in Algorithm 5.1.

Remark B.5. Note that:

(i) Steps 1-b,c,d are sampling from "marginal" posterior of $(p^z_{k^z}, \theta^z_{k^z}, \gamma^z_i)$ for the active components with the auxiliary variables u^z_is being integrated out. Thus, extra caution is needed in dealing with the order of the steps.

(ii) The label switching moves 1-d-i and 1-d-ii are based on Papaspiliopoulos and Roberts (2008), and 1-d-iii is suggested by Hastie *et al.* (2015). All these label switching moves aim to improve numerical convergence.

(iii) Step 3 for DP scale parameter α^z follows Escobar and West (1995). It is different from step 1-a in Algorithm 5.1 due to the unrestricted number of components in the current sampler.

(iv) Steps 4-a-ii and 4-b that update potential components are very similar to steps 1-b and 1-c that update active components—just take $J^z_{k^z}$ as an empty set and draw directly from the prior.

(v) The auxiliary variable u^z_i also appears in step 5 that updates component memberships. The inclusion of auxiliary variables helps determine a finite set of relevant components for each individual i without mechanically truncating the infinite mixture.

C Proofs for Baseline Model

C.1 Posterior Consistency: Random Effects Model

C.1.1 Skills vs Shocks

Proof. **(Proposition 4.7)**

Based on the Schwartz (1965) theorem stated in Lemma 4.6, two sufficient conditions guarantee the posterior consistency: KL requirement and uniformly exponentially consistent tests.

(i) KL requirement

The proposition assumes that the KL property holds for the distribution of λ, i.e. for all $\epsilon > 0$,

$$\Pi^f\left(f \in \mathcal{F} : \int f_0(\lambda) \log \frac{f_0(\lambda)}{f(\lambda)} d\lambda < \epsilon\right) > 0,$$

whose sufficient conditions are stated in Lemmas 4.8 and E.1. On the other hand, the KL requirement is specified on the observed y in order to guarantee that the denominator in equation (4.2) is large enough. In this sense, we need to establish that for all $\epsilon > 0$,

$$\Pi\left(f \in \mathcal{F} : \int f_0(y-u)\,\phi(u) \log \frac{\int f_0(y-u')\,\phi(u')\,du'}{\int f(y-u')\,\phi(u')\,du'} dudy < \epsilon\right) > 0.$$

Let $g(x) = x \log x$, $a(u) = f_0(y-u)\,\phi(u)$, $A = \int a(u)\,du$, $b(u) = f(y-u)\,\phi(u)$, $B = \int b(u)\,du$.

We can rewrite the integral over u as

$$\begin{aligned}
&\int f_0(y-u)\,\phi(u)\log\frac{\int f_0(y-u')\,\phi(u')\,du'}{\int f(y-u')\,\phi(u')\,du'}du = A\cdot\log\frac{A}{B} = B\cdot g\left(\frac{A}{B}\right)\\
=&B\cdot g\left(\int\frac{b(u)}{B}\cdot\frac{f_0(y-u)}{f(y-u)}du\right)\leq\int b(u)\,g\left(\frac{f_0(y-u)}{f(y-u)}\right)du\\
=&\int\phi(u)\,f_0(y-u)\log\frac{f_0(y-u)}{f(y-u)}du, \qquad \text{(C.1)}
\end{aligned}$$

where the inequality is given by Jensen's inequality. Then, further integrating the above expression over y, we have

$$\begin{aligned}
&\int f_0(y-u)\,\phi(u)\log\frac{\int f_0(y-u')\,\phi(u')\,du'}{\int f(y-u')\,\phi(u')\,du'}dudy\leq\int\phi(u)\,f_0(y-u)\log\frac{f_0(y-u)}{f(y-u)}dudy\\
=&\int\phi(u)\,du\cdot\int f_0(\lambda)\log\frac{f_0(\lambda)}{f(\lambda)}d\lambda=\epsilon
\end{aligned}$$

The inequality follows the above expression (C.1), the next equality is given by change of variables, and the last equality is given by the KL property of the distribution of λ.

(ii) Uniformly exponentially consistent tests

(ii-a) When λ is observed

Note that by the Hoeffding's inequality, the uniformly exponentially consistent tests are equivalent to strictly unbiased tests, so we only need to construct a test function $\varphi^\star$ such that

$$\mathbb{E}_{f_0}(\varphi^\star)<\inf_{f\in U^c}\mathbb{E}_f(\varphi^\star).$$

Without loss of generality, let us consider a weak neighborhood defined on $\epsilon>0$ and a bounded continuous function φ ranging from 0 to 1. Then, the corresponding neighborhood is given by

$$U_{\epsilon,\varphi}(f_0)=\left\{f:\ \left|\int\varphi f-\int\varphi f_0\right|<\epsilon\right\}.$$

We can divide the alternative region into two parts[29]

$$U^c_{\epsilon,\varphi}(f_0)=A_1\cup A_2$$

[29]It is legitimate to divide the alternatives into sub-regions. Intuitively, with different alternative sub-regions, the numerator in equation (4.2) is composed of integrals over different domains, and all of them converge to 0.

where

$$A_1 = \left\{ f : \int \varphi f > \int \varphi f_0 + \epsilon \right\},$$
$$A_2 = \left\{ f : \int \varphi f < \int \varphi f_0 - \epsilon \right\}.$$

For A_1, we can choose the test function $\varphi^\star$ to be φ. For A_2, we can choose $\varphi^\star$ to be $1-\varphi$. Then, in either case $A = A_1,\ A_2$, type I error $\mathbb{E}_{f_0}(\varphi^\star) = \int \varphi^\star f_0$, and power $\inf_{f\in A} \mathbb{E}_f(\varphi^\star) \geq \int \varphi^\star f_0 + \epsilon$, hence the tests exist when λ is observed.

(ii-b) When y is observed instead of λ

Define $g(\lambda) = f(\lambda) - f_0(\lambda)$. Then, by definition, $\int g(\lambda)\, d\lambda = 0$ for all g. There are always tests if we observe λ, then for any g, there exists a $\epsilon > 0$ such that

$$\int |g(\lambda)|\, d\lambda > \epsilon. \tag{C.2}$$

The next step is to prove that there are tests when y is observed instead of λ, which is done via proof by contradiction. Suppose there is no test when we only observe y, then there exists a $\tilde{g}$ such that

$$\tilde{h}(y) = \int \tilde{g}(y-u)\,\phi(u)\, du = 0 \text{ for all } y,$$

due to the continuity of $\tilde{h}$. Employing the Fourier transform, we have

$$F_y(\xi) = F_\lambda(\xi) \cdot c_1 \exp\left(-c_2\xi^2\right) = 0 \text{ for all } \xi.$$

Since $c_1 \exp\left(-c_2\xi^2\right) \neq 0$, then

$$F_\lambda(\xi) = 0 \text{ for all } \xi.$$

Finally, the inverse Fourier transform leads to

$$\tilde{g}(\lambda) = 0 \text{ for all } \lambda,$$

which contradicts equation (C.2). Therefore, there are also tests when y is observed instead of λ.

Combining (i) and (ii-b), f achieves posterior consistency even when we only observe y. □

C.1.2 Unknown Shocks Sizes

Proof. **(Proposition 4.9)**

(i) KL requirement

Based on the observed sufficient statistics $\hat{\lambda} = \frac{1}{T}\sum_{t=1}^{T} y_{it}$ with corresponding errors $\hat{u} = \frac{1}{T}\sum_{t=1}^{T} u_{it}$,

the KL requirement can be written as follows: for all $\epsilon > 0$,

$$\Pi\left(\begin{array}{l} f \in \mathcal{F},\ \sigma^2 \in \mathbb{R}^+ : \\ \int f_0\left(\hat{\lambda}-\hat{u}\right)\phi\left(\hat{u};\ 0, \frac{\sigma_0^2}{T}\right)\log\frac{\int f_0\left(\hat{\lambda}-\hat{u}'\right)\phi\left(\hat{u}';\ 0, \frac{\sigma_0^2}{T}\right)d\hat{u}'}{\int f\left(\hat{\lambda}-\hat{u}'\right)\phi\left(\hat{u}';\ 0, \frac{\sigma^2}{T}\right)d\hat{u}'}d\hat{u}d\hat{\lambda} < \epsilon \end{array}\right) > 0.$$

Under the prior specification together with hyperparameters specified in Appendix B.1, the integral is bounded with probability one. Following the dominated convergence theorem,

$$\begin{aligned} &\lim_{\sigma^2\to\sigma_0^2}\int f_0\left(\hat{\lambda}-\hat{u}\right)\phi\left(\hat{u};\ 0, \frac{\sigma_0^2}{T}\right)\log\frac{\int f_0\left(\hat{\lambda}-\hat{u}'\right)\phi\left(\hat{u}';\ 0, \frac{\sigma_0^2}{T}\right)d\hat{u}'}{\int f\left(\hat{\lambda}-\hat{u}'\right)\phi\left(\hat{u}';\ 0, \frac{\sigma^2}{T}\right)d\hat{u}'}d\hat{u}d\hat{\lambda} \\ =&\int f_0\left(\hat{\lambda}-\hat{u}\right)\phi\left(\hat{u};\ 0, \frac{\sigma_0^2}{T}\right)\log\frac{\int f_0\left(\hat{\lambda}-\hat{u}'\right)\phi\left(\hat{u}';\ 0, \frac{\sigma_0^2}{T}\right)d\hat{u}'}{\int f\left(\hat{\lambda}-\hat{u}'\right)\phi\left(\hat{u}';\ 0, \frac{\sigma^2}{T}\right)d\hat{u}'}d\hat{u}d\hat{\lambda}, \end{aligned}$$

where the upper bound of the right hand side can be characterized by the KL property of the distribution of λ as in the proof of Proposition 4.7 part (i). The sufficient conditions of the KL property of the distribution of λ are stated in Lemmas 4.8 and E.1.

(ii) Uniformly exponentially consistent tests

The alternative region can be split into the following two parts:

(ii-a) $\left|\sigma^2-\sigma_0^2\right| > \Delta$

Orthogonal forward differencing yields $\tilde{y}_{it} \sim N\left(0, \sigma_0^2\right)$. Then, as $N \to \infty$,

$$\frac{\frac{1}{N(T-1)}\sum_{i=1}^{N}\sum_{t=1}^{T-1}\left(\tilde{y}_{it}\right)^2}{\sigma_0^2} \sim \chi^2_{N(T-1)} \xrightarrow{d} N\left(1, \frac{2}{N\left(T-1\right)}\right).$$

Note that for a generic variable $x \sim N\left(0,1\right)$, for $x^* > 0$,

$$\mathbb{P}\left(x > x^*\right) \leq \frac{\phi\left(x^*\right)}{x^*}. \tag{C.3}$$

Then, we can directly construct the following test function

$$\boldsymbol{\varphi_N}\left(\tilde{y}_{1:N,1:T-1}\right) = \begin{cases} \mathbf{1}\left(\frac{\frac{1}{N(T-1)}\sum_{i=1}^{N}\sum_{t=1}^{T-1}(\tilde{y}_{it})^2}{\sigma_0^2} < 1-\frac{\Delta}{2\sigma_0^2}\right), & for\ \sigma^2 < \sigma_0^2-\Delta, \\ \mathbf{1}\left(\frac{\frac{1}{N(T-1)}\sum_{i=1}^{N}\sum_{t=1}^{T-1}(\tilde{y}_{it})^2}{\sigma_0^2} > 1+\frac{\Delta}{2\sigma_0^2}\right), & for\ \sigma^2 > \sigma_0^2+\Delta, \end{cases}$$

which satisfies the requirements (4.1) for the uniformly exponentially consistent tests.

(ii-b) $\left|\sigma^2-\sigma_0^2\right| < \Delta,\ f \in U^c_{\epsilon,\Phi}\left(f_0\right)$

Without loss of generality, let $\Phi = \{\varphi\}$ be a singleton and $\varphi^\star$ be the test function that distin-

guishes $f = f_0$ versus $f \in U^c_{\epsilon,\varphi}(f_0)$ when σ_0^2 is known. Then, we can express the difference between $\mathbb{E}_f(\varphi^\star)$ and $\mathbb{E}_{f_0}(\varphi^\star)$ as

$$\begin{aligned}
&\int \varphi^\star\left(\hat{\lambda}\right) f\left(\hat{\lambda}-\hat{u}\right)\phi\left(\hat{u};\,0,\frac{\sigma^2}{T}\right)d\hat{u}d\hat{\lambda} - \int \varphi^\star\left(\hat{\lambda}\right) f_0\left(\hat{\lambda}-\hat{u}\right)\phi\left(\hat{u};\,0,\frac{\sigma_0^2}{T}\right)d\hat{u}d\hat{\lambda}\\
>&\int \varphi^\star\left(\hat{\lambda}\right)\left(f\left(\hat{\lambda}-\hat{u}\right)-f_0\left(\hat{\lambda}-\hat{u}\right)\right)\phi\left(\hat{u};\,0,\frac{\sigma_0^2}{T}\right)d\hat{u}d\hat{\lambda}\\
&-\left|\int \varphi^\star\left(\hat{\lambda}\right) f\left(\hat{\lambda}-\hat{u}\right)\left(\phi\left(\hat{u};\,0,\frac{\sigma^2}{T}\right)-\phi\left(\hat{u};\,0,\frac{\sigma_0^2}{T}\right)\right)d\hat{u}d\hat{\lambda}\right|. \qquad \text{(C.4)}
\end{aligned}$$

Since $\varphi^\star$ is the test function when σ_0^2 is known, the first term

$$\int \varphi^\star\left(\hat{\lambda}\right)\left(f\left(\hat{\lambda}-\hat{u}\right)-f_0\left(\hat{\lambda}-\hat{u}\right)\right)\phi\left(\hat{u};\,0,\frac{\sigma_0^2}{T}\right)d\hat{u}d\hat{\lambda} > \epsilon. \qquad \text{(C.5)}$$

For the second term,

$$\begin{aligned}
&\left|\int \varphi^\star\left(\hat{\lambda}\right) f\left(\hat{\lambda}-\hat{u}\right)\left(\phi\left(\hat{u};\,0,\frac{\sigma^2}{T}\right)-\phi\left(\hat{u};\,0,\frac{\sigma_0^2}{T}\right)\right)d\hat{u}d\hat{\lambda}\right|\\
\leq&\int \varphi^\star\left(\hat{\lambda}\right) f\left(\hat{\lambda}-\hat{u}\right)\left|\phi\left(\hat{u};\,0,\frac{\sigma^2}{T}\right)-\phi\left(\hat{u};\,0,\frac{\sigma_0^2}{T}\right)\right|d\hat{u}d\hat{\lambda}\\
\leq&\int \left|\phi\left(\hat{u};\,0,\frac{\sigma^2}{T}\right)-\phi\left(\hat{u};\,0,\frac{\sigma_0^2}{T}\right)\right|d\hat{u}\\
\leq&\sqrt{\frac{\sigma_0^2}{\sigma^2}-1-\ln\frac{\sigma_0^2}{\sigma^2}}. \qquad \text{(C.6)}
\end{aligned}$$

The second inequality is given by the fact that $\varphi^\star\left(\hat{\lambda}\right) \in [0,1]$. The last inequality follows Pinsker's inequality that bounds the total variation distance by the KL divergence, which has an explicit form for normal distributions

$$d_{KL}\left(\phi\left(\hat{u};\,0,\frac{\sigma_0^2}{T}\right),\phi\left(\hat{u};\,0,\frac{\sigma^2}{T}\right)\right) = \frac{1}{2}\left(\frac{\sigma_0^2}{\sigma^2}-1-\ln\frac{\sigma_0^2}{\sigma^2}\right).$$

We can choose $\Delta > 0$ such that for any $\left|\sigma^2-\sigma_0^2\right| < \Delta$,

$$\sqrt{\frac{\sigma_0^2}{\sigma^2}-1-\ln\frac{\sigma_0^2}{\sigma^2}} < \frac{\epsilon}{2}.$$

Plugging expressions (C.5) and (C.6) into (C.4), we obtain

$$\int \varphi^\star\left(\hat{\lambda}\right) f\left(\hat{\lambda}-\hat{u}\right)\phi\left(\hat{u};\,0,\frac{\sigma^2}{T}\right)d\hat{u}d\hat{\lambda} - \int \varphi^\star\left(\hat{\lambda}\right) f_0\left(\hat{\lambda}-\hat{u}\right)\phi\left(\hat{u};\,0,\frac{\sigma_0^2}{T}\right)d\hat{u}d\hat{\lambda} > \epsilon - \frac{\epsilon}{2} = \frac{\epsilon}{2},$$

so $\varphi^\star$ is the test function with respect to the alternative sub-region $\left\{\left|\sigma^2-\sigma_0^2\right| < \Delta,\ f \in U^c_{\epsilon,\Phi}(f_0)\right\}$.

□

C.1.3 Lagged Dependent Variables

Proof. **(Proposition 4.11)**

(i) KL requirement

Define the sufficient statistics $\hat{\lambda}(\beta) = \frac{1}{T}\sum_{t=1}^{T} y_{it} - \beta y_{i,t-1}$ with corresponding errors $\hat{u} = \frac{1}{T}\sum_{t=1}^{T} u_{it}$. The KL requirement is satisfied as long as for all $\epsilon > 0$,

$$\Pi\left(\begin{array}{l} f \in \mathcal{F},\ \left(\beta, \sigma^2\right) \in \mathbb{R} \times \mathbb{R}^+ : \\ \int f_0\left(\hat{\lambda}(\beta_0) - \hat{u}\right) \phi\left(\hat{u};\ 0, \frac{\sigma_0^2}{T}\right) \log \frac{\int f_0\left(\hat{\lambda}(\beta_0) - \hat{u}'\right) \phi\left(\hat{u}';\ 0, \frac{\sigma_0^2}{T}\right) d\hat{u}'}{\int f\left(\hat{\lambda}(\beta) - \hat{u}'\right) \phi\left(\hat{u}';\ 0, \frac{\sigma^2}{T}\right) d\hat{u}'} d\hat{u} d\hat{\lambda} < \epsilon \end{array}\right) > 0.$$

Similar to the previous case, the dominated convergence theorem and the KL property of the distribution of λ complete the proof.

(ii) Uniformly exponentially consistent tests

The alternative region can be split into the following two parts:

(ii-a) $|\beta - \beta_0| > \Delta$ or $\left|\sigma^2 - \sigma_0^2\right| > \Delta'$

Orthogonal forward differencing yields $\tilde{y}_{it} = \beta \tilde{y}_{i,t-1} + \tilde{u}_{it},\ \tilde{u}_{it} \sim N\left(0, \sigma_0^2\right)$. Then, as $N \to \infty$,

$$\hat{\beta}_{OLS} = \left(\sum_{i=1}^{N}\sum_{t=1}^{T-1} \left(\tilde{y}_{i,t-1}\right)^2\right)^{-1} \left(\sum_{i=1}^{N}\sum_{t=1}^{T-1} \tilde{y}_{i,t-1}\tilde{y}_{it}\right) \xrightarrow{d} N\left(\beta_0, \frac{\sigma_0^2}{N\sum_{t=1}^{T-1}\mathbb{E}\left(\tilde{y}_{i,t-1}\right)^2}\right)$$

$$\frac{\frac{1}{N(T-1)}\sum_{i=1}^{N}\sum_{t=1}^{T-1}\left(\tilde{y}_{it} - \hat{\beta}_{OLS}\tilde{y}_{i,t-1}\right)^2}{\sigma_0^2} \sim \chi^2_{N(T-1)-1} \xrightarrow{d} N\left(1, \frac{2}{N(T-1)-1}\right).$$

Since the upper tail of a normal distribution is bounded as in expression (C.3), we can directly construct the following test function

$$\boldsymbol{\varphi_N} = 1 - (1 - \boldsymbol{\varphi_{N,\beta}})\left(1 - \boldsymbol{\varphi_{N,\sigma^2}}\right),$$

where

$$\boldsymbol{\varphi_{N,\beta}}\left(\tilde{y}_{1:N,1:T-1}\right) = \begin{cases} \mathbf{1}\left(\hat{\beta}_{OLS} < \beta_0 - \frac{\Delta}{2}\right), & for\ \beta < \beta_0 - \Delta, \\ \mathbf{1}\left(\hat{\beta}_{OLS} > \beta_0 + \frac{\Delta}{2}\right), & for\ \beta > \beta_0 + \Delta, \end{cases}$$

$$\boldsymbol{\varphi_{N,\sigma^2}}\left(\tilde{y}_{1:N,1:T-1}\right) = \begin{cases} \mathbf{1}\left(\frac{\frac{1}{N(T-1)}\sum_{i=1}^{N}\sum_{t=1}^{T-1}\left(\tilde{y}_{it} - \hat{\beta}_{OLS}\tilde{y}_{i,t-1}\right)^2}{\sigma_0^2} < 1 - \frac{\Delta'}{2\sigma_0^2}\right), & for\ \sigma^2 < \sigma_0^2 - \Delta', \\ \mathbf{1}\left(\frac{\frac{1}{N(T-1)}\sum_{i=1}^{N}\sum_{t=1}^{T-1}\left(\tilde{y}_{it} - \hat{\beta}_{OLS}\tilde{y}_{i,t-1}\right)^2}{\sigma_0^2} > 1 + \frac{\Delta'}{2\sigma_0^2}\right), & for\ \sigma^2 > \sigma_0^2 + \Delta', \end{cases}$$

which satisfies the requirements (4.1) for the uniformly exponentially consistent tests.

(ii-b) $|\beta - \beta_0| < \Delta,\ \left|\sigma^2 - \sigma_0^2\right| < \Delta',\ f \in U^c_{\epsilon,\Phi}(f_0)$

The following proof is analogous to the proofs of Proposition 3.3 in Amewou-Atisso *et al.* (2003) except the inclusion of shocks u_{it}s in the current setup, which prohibits direct inference of λ_i. Without loss of generality, let $\Phi = \{\varphi\}$ and $\varphi^\star(\mathring{y})$ be the corresponding test function on $\mathring{y} = y_{i1} - \beta_0 y_{i0} = \lambda_i + u_{i1}$ when β_0 and σ_0^2 are known. Then, we can construct a uniformly continuous test function

$$\varphi^{\star\star}(\mathring{y}) = \begin{cases} \varphi^\star(\mathring{y}), & if\ |\mathring{y}| < M_1, \\ 1, & if\ |\mathring{y}| > M_2, \\ \max\left\{\varphi^\star(\mathring{y}),\ \varphi^\star(M_1) + \frac{1-\varphi^\star(M_1)}{M_2 - M_1}(\mathring{y} - M_1)\right\}, & if\ \mathring{y} \in [M_1, M_2], \\ \max\left\{\varphi^\star(\mathring{y}),\ 1 + \frac{\varphi^\star(-M_1)-1}{M_2-M_1}(\mathring{y} + M_2)\right\} & if\ \mathring{y} \in [-M_2, -M_1], \end{cases}$$

where M_1 is chosen such that

$$\int_{|\mathring{y}|>M_1} f_0(\mathring{y} - u)\,\phi\left(u;\ 0, \sigma_0^2\right) du dy_1 < \frac{\epsilon}{4}.$$

Then,

$$\int \varphi^{\star\star}(\mathring{y})\, f(\mathring{y} - u)\,\phi\left(u;\ 0, \sigma_0^2\right) du dy_1 - \int \varphi^{\star\star}(\mathring{y})\, f_0(\mathring{y} - u)\,\phi\left(u;\ 0, \sigma_0^2\right) du dy_1 > \frac{3}{4}\epsilon. \tag{C.7}$$

Due to uniform continuity, given $\epsilon > 0$, there exists $\delta > 0$ such that $|\varphi^{\star\star}(\mathring{y}') - \varphi^{\star\star}(\mathring{y})| < \epsilon/4$ for any $|\mathring{y}' - \mathring{y}| < \delta$. As y_{i0} is compacted supported, we can choose Δ such that $|(\beta - \beta_0) y_{i0}| < \delta$.

Let y_1 be a generic variable representing y_{i1}. Define the test function for the non-i.i.d. case to be $\varphi_i(y_1) = \varphi^{\star\star}(y_1 - \beta_0 y_{i0})$. Then, the difference between $\mathbb{E}_f(\varphi_i)$ and $\mathbb{E}_{f_0}(\varphi_i)$ is

$$\begin{aligned} &\int \varphi_i(y_1)\, f(y_1 - \beta y_{i0} - u)\,\phi\left(u;\ 0, \sigma^2\right) du dy_1 - \int \varphi_i(y_1)\, f_0(y_1 - \beta_0 y_{i0} - u)\,\phi\left(u;\ 0, \sigma_0^2\right) du dy_1 \\ > &\int \varphi_i(y_1)\left(f(y_1 - \beta_0 y_{i0} - u) - f_0(y_1 - \beta y_{i0} - u)\right)\phi\left(u;\ 0, \sigma_0^2\right) du dy_1 \\ &+ \int \varphi_i(y_1)\left(f(y_1 - \beta y_{i0} - u) - f(y_1 - \beta_0 y_{i0} - u)\right)\phi\left(u;\ 0, \sigma_0^2\right) du dy_1 \\ &- \left|\int \varphi_i(y_1)\, f(y_1 - \beta y_{i0} - u)\left(\phi\left(u;\ 0, \sigma^2\right) - \phi\left(u;\ 0, \sigma_0^2\right)\right) du dy_1\right|. \end{aligned}$$

From expression (C.7), the first term is bounded below by $3\epsilon/4$. Similar to the proof of Proposition 4.9 part (ii-b), the third term is bounded above by $\epsilon/4$. For the second term, note that for any δ,

$$\int \varphi^{\star\star}(y_1 - \delta)\, f(y_1 - \delta - u)\, dy_1 = \int \varphi^{\star\star}(y_1)\, f(y_1 - u)\, dy_1$$

Then,

$$\begin{aligned}
&\int \varphi_i(y_1)\left(f(y_1-\beta y_{i0}-u)-f(y_1-\beta_0 y_{i0}-u)\right)dy_1 \\
=&\int \varphi^{\star\star}(y_1+(\beta-\beta_0)y_{i0})\,f(y_1-u)\,dy_1-\int \varphi^{\star\star}(y_1)\,f(y_1-u)\,dy_1 \\
\geq&-\int \left|\varphi^{\star\star}(y_1+(\beta-\beta_0)y_{i0})-\varphi^{\star\star}(y_1)\right| f(y_1-u)\,dy_1 \\
\geq&-\frac{\epsilon}{4}
\end{aligned}$$

where the last inequality is given by the uniform continuity of $\varphi^{\star\star}$. Hence, $\mathbb{E}_f(\varphi_i)-\mathbb{E}_{f_0}(\varphi_i)>\epsilon/4$, and $\{\varphi_i\}$ constitutes the tests with respect to the alternative sub-region $\left\{|\beta-\beta_0|<\Delta,\ \left|\sigma^2-\sigma_0^2\right|<\Delta',\ f\in U^c_{\epsilon,\Phi}(f_0)\right\}$. □

C.2 Posterior Consistency: Correlated Random Effects Model

Recall that h, f, and q are the joint, conditional, and marginal densities, respectively. In addition,

$$h_0(\lambda,c)=f_0(\lambda|c)\cdot q_0(c),\quad h(\lambda,c)=f(\lambda|c)\cdot q_0(c).$$

Proof. **(Proposition 4.15)**

(i) KL requirement

Define the sufficient statistics $\hat{\lambda}(\beta)=\frac{1}{T}\sum_{t=1}^T y_{it}-\beta y_{i,t-1}$ with corresponding errors $\hat{u}=\frac{1}{T}\sum_{t=1}^T u_{it}$. Considering joint density characterization, the observations are i.i.d. across i in the correlated random effects setup. The KL requirement can be specified as follows: for all $\epsilon>0$,

$$\Pi\left(\begin{array}{l} f\in\mathcal{F},\ (\beta,\sigma^2)\in\mathbb{R}\times\mathbb{R}^+: \\ \int h_0\left(\hat{\lambda}(\beta_0)-\hat{u},\ y_0\right)\phi\left(\hat{u};\ 0,\frac{\sigma_0^2}{T}\right)\log\dfrac{\int h_0\left(\hat{\lambda}(\beta_0)-\hat{u}',\ y_0\right)\phi\left(\hat{u}';\ 0,\frac{\sigma_0^2}{T}\right)d\hat{u}'}{\int h\left(\hat{\lambda}(\beta)-\hat{u}',\ y_0\right)\phi\left(\hat{u}';\ 0,\frac{\sigma^2}{T}\right)d\hat{u}'}d\hat{u}d\hat{\lambda}dy_0<\epsilon \end{array}\right)>0.$$

The rest of the proof is similar to the previous cases employing the dominated convergence theorem and the KL property of the joint distribution of (λ,y_0) with sufficient conditions stated in Assumption 4.14.

(ii) Uniformly exponentially consistent tests

It follows the proof of Proposition 4.11 part (ii) except that in case $|\beta-\beta_0|<\Delta,\ \left|\sigma^2-\sigma_0^2\right|<\Delta',\ f\in U^c_{\epsilon,\Phi}(f_0)$, the test function φ is defined on (y_1,y_0) that distinguishes the true h_0 from alternative h. □

C.3 Density Forecasts

Proof. **(Proposition 4.16)**

(i) Random Effects: Result 1

In this part, I am going to prove that for any i and any $U_{\epsilon,\Phi}\left(f_{i,T+1}^{oracle}\right)$, as $N \to \infty$,

$$\mathbb{P}\left(f_{i,T+1}^{cond} \in U_{\epsilon,\Phi}\left(f_{i,T+1}^{oracle}\right) \middle| y_{1:N,0:T}\right) \to 1,\ a.s.$$

This is equivalent to proving that for any bounded continuous function φ,

$$\mathbb{P}\left(f \in \mathcal{F}:\ \left|\int \varphi(y) f_{i,T+1}^{cond}\left(y|\beta,\sigma^2,f,y_{i,0:T}\right)dy - \int \varphi(y) f_{i,T+1}^{oracle}(y)\,dy\right| < \epsilon \middle| y_{1:N,0:T}\right) \to 1,\ a.s.$$

where

$$\begin{aligned}
&\left|\int \varphi(y) f_{i,T+1}^{cond}\left(y|\beta,\sigma^2,f,y_{i,0:T}\right)dy - \int \varphi(y) f_{i,T+1}^{oracle}(y)\,dy\right| \\
=&\left|\int \varphi(y)\,\phi\left(y;\ \beta y_{iT}+\lambda_i,\sigma^2\right) p\left(\lambda_i\left|\beta,\sigma^2,f,y_{i,0:T}\right.\right) d\lambda_i dy\right. \\
&\left.-\int \varphi(y)\,\phi\left(y;\ \beta_0 y_{iT}+\lambda_i,\sigma_0^2\right) p\left(\lambda_i\left|\beta_0,\sigma_0^2,f_0,y_{i,0:T}\right.\right) d\lambda_i dy\right| \\
=&\left|\frac{\int \varphi(y)\,\phi\left(y;\ \beta y_{iT}+\lambda_i,\sigma^2\right)\prod_t p\left(y_{it}\left|\lambda_i,\beta,\sigma^2,y_{i,t-1}\right.\right) f(\lambda_i)\,d\lambda_i dy}{\int \prod_t p\left(y_{it}\left|\lambda_i,\beta,\sigma^2,y_{i,t-1}\right.\right) f(\lambda_i)\,d\lambda_i}\right. \\
&\left.-\frac{\int \varphi(y)\,\phi\left(y;\ \beta_0 y_{iT}+\lambda_i,\sigma_0^2\right)\prod_t p\left(y_{it}\left|\lambda_i,\beta_0,\sigma_0^2,y_{i,t-1}\right.\right) f_0(\lambda_i)\,d\lambda_i dy}{\int \prod_t p\left(y_{it}\left|\lambda_i,\beta_0,\sigma_0^2,y_{i,t-1}\right.\right) f_0(\lambda_i)\,d\lambda_i}\right|.
\end{aligned}$$

The last equality is given by plugging in

$$p\left(\lambda_i\left|\beta,\sigma^2,f,y_{i,0:T}\right.\right) = \frac{\prod_t p\left(y_{it}\left|\lambda_i,\beta,\sigma^2,y_{i,t-1}\right.\right) f(\lambda_i)}{\int \prod_t p\left(y_{it}\left|\lambda_i',\beta,\sigma^2,y_{i,t-1}\right.\right) f(\lambda_i')\,d\lambda_i'}.$$

Set

$$\begin{aligned}
A &= \int \prod_t p\left(y_{it}\left|\lambda_i,\beta,\sigma^2,y_{i,t-1}\right.\right) d\lambda_i, \\
B &= \int \varphi(y)\,\phi\left(y;\ \beta y_{iT}+\lambda_i,\sigma^2\right)\prod_t p\left(y_{it}\left|\lambda_i,\beta,\sigma^2,y_{i,t-1}\right.\right) d\lambda_i dy.
\end{aligned}$$

with A_0 and B_0 being the counterparts for the oracle predictor. Then, we want to make sure the following expression is arbitrarily small,

$$\left|\frac{B}{A}-\frac{B_0}{A_0}\right| \le \frac{|B_0|\,|A-A_0|}{|A_0|\,|A|}+\frac{|B-B_0|}{|A|},$$

and it is sufficient to establish the following four statements.

(a) $|A - A_0| < \epsilon'$

$$
\begin{aligned}
&|A - A_0| \\
\leq &\left| \int \prod_t p\left(y_{it} \left| \lambda_i, \beta_0, \sigma_0^2, y_{i,t-1}\right.\right) \left(f\left(\lambda_i\right) - f_0\left(\lambda_i\right)\right) d\lambda_i \right| \\
&+ \left| \int \left(\prod_t p\left(y_{it} \left| \lambda_i, \beta, \sigma^2, y_{i,t-1}\right.\right) - \prod_t p\left(y_{it} \left| \lambda_i, \beta_0, \sigma_0^2, y_{i,t-1}\right.\right) \right) f_0\left(\lambda_i\right) d\lambda_i \right|
\end{aligned}
$$

The first term is less than $\epsilon'/2$ with probability one due to the posterior consistency of f and that

$$
\prod_t p\left(y_{it} \left| \lambda_i, \beta_0, \sigma_0^2, y_{i,t-1}\right.\right) = C\left(\beta_0, \sigma_0^2, y_{i,0:T}\right) \phi\left(\lambda_i;\ \frac{1}{T}\sum_T \left(y_{it} - \beta_0 y_{i,t-1}\right), \frac{\sigma_0^2}{T}\right) \tag{C.8}
$$

is a bounded continuous function in λ_i, with $C\left(\beta_0, \sigma_0^2, y_{i,0:T}\right)$ being

$$
C\left(\beta_0, \sigma_0^2, y_{i,0:T}\right) = \frac{1}{\sqrt{T}\left(2\pi\sigma_0^2\right)^{\frac{T-1}{2}}} \exp\left(-\frac{\sum_t \left(y_{it} - \beta_0 y_{i,t-1}\right)^2 - \frac{1}{T}\left(\sum_T \left(y_{it} - \beta_0 y_{i,t-1}\right)\right)^2}{2\sigma_0^2} \right).
$$

For the second term,

$$
\begin{aligned}
&\left| \int \left(\prod_t p\left(y_{it} \left| \lambda_i, \beta, \sigma^2, y_{i,t-1}\right.\right) - \prod_t p\left(y_{it} \left| \lambda_i, \beta_0, \sigma_0^2, y_{i,t-1}\right.\right) \right) f_0\left(\lambda_i\right) d\lambda_i \right| \\
\leq &M \int \left| \prod_t p\left(y_{it} \left| \lambda_i, \beta, \sigma^2, y_{i,t-1}\right.\right) - \prod_t p\left(y_{it} \left| \lambda_i, \beta_0, \sigma_0^2, y_{i,t-1}\right.\right) \right| d\lambda_i \\
\leq &MC\left(\beta_0, \sigma_0^2, y_{i,0:T}\right) \int \left| \phi\left(\lambda_i;\ \frac{1}{T}\sum_T \left(y_{it} - \beta y_{i,t-1}\right), \frac{\sigma^2}{T}\right) - \phi\left(\lambda_i;\ \frac{1}{T}\sum_T \left(y_{it} - \beta_0 y_{i,t-1}\right), \frac{\sigma_0^2}{T}\right) \right| d\lambda_i \\
&+ M\left|C\left(\beta, \sigma^2, y_{i,0:T}\right) - C\left(\beta_0, \sigma_0^2, y_{i,0:T}\right)\right| \int \phi\left(\lambda_i;\ \frac{1}{T}\sum_T \left(y_{it} - \beta y_{i,t-1}\right), \frac{\sigma^2}{T}\right) d\lambda_i.
\end{aligned} \tag{C.9}
$$

where the last inequality is given by rewriting $\prod_t p\left(y_{it} \left| \lambda_i, \beta, \sigma^2, y_{i,t-1}\right.\right)$ as a distribution of λ_i (equation C.8). Following Pinsker's inequality that bounds the total variation distance by the KL

divergence,

$$\begin{aligned}
&\int \left| \phi\left(\lambda_i;\ \frac{1}{T}\sum_T \left(y_{it}-\beta y_{i,t-1}\right), \frac{\sigma^2}{T}\right) - \phi\left(\lambda_i;\ \frac{1}{T}\sum_T \left(y_{it}-\beta_0 y_{i,t-1}\right), \frac{\sigma_0^2}{T}\right)\right| d\lambda_i \\
\leq &\sqrt{2 d_{KL}\left(\phi\left(\lambda_i;\ \frac{1}{T}\sum_T \left(y_{it}-\beta_0 y_{i,t-1}\right), \frac{\sigma_0^2}{T}\right),\ \phi\left(\lambda_i;\ \frac{1}{T}\sum_T \left(y_{it}-\beta y_{i,t-1}\right), \frac{\sigma^2}{T}\right)\right)} \\
\leq &\sqrt{\frac{\sigma_0^2}{\sigma^2} - 1 - \ln\frac{\sigma_0^2}{\sigma^2} + \frac{(\beta-\beta_0)^2\left(\sum_t y_{i,t-1}\right)^2}{T\sigma^2}}. \qquad \text{(C.10)}
\end{aligned}$$

As (β, σ^2) enjoy posterior consistency, both $\left|C\left(\beta, \sigma^2, y_{i,0:T}\right) - C\left(\beta_0, \sigma_0^2, y_{i,0:T}\right)\right|$ in expression (C.9) and $\sqrt{\frac{\sigma_0^2}{\sigma^2} - 1 - \ln\frac{\sigma_0^2}{\sigma^2} + \frac{(\beta-\beta_0)^2\left(\sum_t y_{i,t-1}\right)^2}{T\sigma^2}}$ in expression (C.10) can be arbitrarily small. Therefore, the second term is less than $\epsilon'/2$ with probability one.

(b) $|B - B_0| < \epsilon'$

$$\begin{aligned}
&|B - B_0| \\
\leq &\left| \int \varphi(y)\, \phi\left(y;\ \beta_0 y_{iT} + \lambda_i, \sigma_0^2\right) \prod_t p\left(y_{it}\left|\lambda_i, \beta_0, \sigma_0^2, y_{i,t-1}\right.\right) \left(f(\lambda_i) - f_0(\lambda_i)\right) d\lambda_i dy \right| \\
&+ \left| \int \varphi(y) \left(\begin{array}{l} \phi\left(y;\ \beta y_{iT} + \lambda_i, \sigma^2\right) \prod_t p\left(y_{it}\left|\lambda_i, \beta, \sigma^2, y_{i,t-1}\right.\right) \\ \quad - \phi\left(y;\ \beta_0 y_{iT} + \lambda_i, \sigma_0^2\right) \prod_t p\left(y_{it}\left|\lambda_i, \beta_0, \sigma_0^2, y_{i,t-1}\right.\right) \end{array} \right) f_0(\lambda_i)\, d\lambda_i dy \right|
\end{aligned}$$

Similar to (a), the first term is small due to the posterior consistency of f, while Pinsker's inequality together with the posterior consistency of (β, σ^2) ensure a small second term.

(c) There exists $\underline{A} > 0$ such that $|A_0| > \underline{A}$.

$$\begin{aligned}
A_0 &= \int \prod_t p\left(y_{it}\left|\lambda_i, \beta_0, \sigma_0^2, y_{i,t-1}\right.\right) f_0(\lambda_i)\, d\lambda_i \\
&= C\left(\beta_0, \sigma_0^2, y_{i,0:T}\right) \int \phi\left(\lambda_i;\ \frac{1}{T}\sum_T \left(y_{it}-\beta_0 y_{i,t-1}\right), \frac{\sigma_0^2}{T}\right) f_0(\lambda_i)\, d\lambda_i
\end{aligned}$$

Since $\phi\left(\lambda_i;\ \frac{1}{T}\sum_T \left(y_{it}-\beta_0 y_{i,t-1}\right), \frac{\sigma_0^2}{T}\right)$ and $f_0(\lambda_i)$ share the same support on $\mathbb{R}$, the integral is bounded below by some positive $\underline{A}$. Moreover, we have $|A - A_0| < \epsilon'$ from (a), then $|A| > |A_0| - \epsilon' > \underline{A} - \epsilon'$. Therefore, both $|A_0|$ and $|A|$ are bounded below.

(d) $|B_0| < \infty$

$$\begin{aligned}
|B_0| &= \left| \int \varphi(y)\, \phi\left(y;\ \beta_0 y_{iT} + \lambda_i, \sigma_0^2\right) \prod_t p\left(y_{it} \left| \lambda_i, \beta_0, \sigma_0^2, y_{i,t-1}\right.\right) f_0(\lambda_i)\, d\lambda_i dy \right| \\
&\leq M_\varphi \cdot \frac{1}{\left(2\pi\sigma_0^2\right)^{\frac{T}{2}}} \cdot \left| \int \phi\left(y;\ \beta_0 y_{iT} + \lambda_i, \sigma_0^2\right) f_0(\lambda_i)\, d\lambda_i dy \right| \\
&= M_\varphi \cdot \frac{1}{\left(2\pi\sigma_0^2\right)^{\frac{T}{2}}}
\end{aligned}$$

(ii) Random Effects: Result 2

Now the goal is to prove that for any i, any y, and any $\epsilon > 0$, as $N \to \infty$,

$$\left| f_{i,T+1}^{sp}(y) - f_{i,T+1}^{oracle}(y) \right| < \epsilon,\ a.s.$$

where

$$\begin{aligned}
&\left| f_{i,T+1}^{sp}(y) - f_{i,T+1}^{oracle}(y) \right| \\
=&\left| \int \phi\left(y;\ \beta y_{iT} + \lambda_i, \sigma^2\right) p\left(\lambda_i \left| \beta, \sigma^2, f, y_{i,0:T}\right.\right) d\Pi\left(\beta, \sigma^2, f \left| y_{1:N,0:T}\right.\right) d\lambda_i d\beta d\sigma^2 df \right. \\
&\left. - \int \phi\left(y;\ \beta_0 y_{iT} + \lambda_i, \sigma_0^2\right) p\left(\lambda_i \left| \beta_0, \sigma_0^2, f_0, y_{i,0:T}\right.\right) d\lambda_i \right| \\
=&\left| \int \frac{\int \phi\left(y;\ \beta y_{iT} + \lambda_i, \sigma^2\right) \prod_t p\left(y_{it} \left| \lambda_i, \beta, \sigma^2, y_{i,t-1}\right.\right) f(\lambda_i)\, d\lambda_i dy}{\int \prod_t p\left(y_{it} \left| \lambda_i, \beta, \sigma^2, y_{i,t-1}\right.\right) f(\lambda_i)\, d\lambda_i} d\Pi\left(\beta, \sigma^2, f \left| y_{1:N,0:T}\right.\right) d\beta d\sigma^2 df \right. \\
&\left. - \frac{\int \phi\left(y;\ \beta_0 y_{iT} + \lambda_i, \sigma_0^2\right) \prod_t p\left(y_{it} \left| \lambda_i, \beta_0, \sigma_0^2, y_{i,t-1}\right.\right) f_0(\lambda_i)\, d\lambda_i dy}{\int \prod_t p\left(y_{it} \left| \lambda_i, \beta_0, \sigma_0^2, y_{i,t-1}\right.\right) f_0(\lambda_i)\, d\lambda_i} \right| \\
\leq&\int \left| \frac{\int \phi\left(y;\ \beta y_{iT} + \lambda_i, \sigma^2\right) \prod_t p\left(y_{it} \left| \lambda_i, \beta, \sigma^2, y_{i,t-1}\right.\right) f(\lambda_i)\, d\lambda_i dy}{\int \prod_t p\left(y_{it} \left| \lambda_i, \beta, \sigma^2, y_{i,t-1}\right.\right) f(\lambda_i)\, d\lambda_i} \right. \\
&\left. - \frac{\int \phi\left(y;\ \beta_0 y_{iT} + \lambda_i, \sigma_0^2\right) \prod_t p\left(y_{it} \left| \lambda_i, \beta_0, \sigma_0^2, y_{i,t-1}\right.\right) f_0(\lambda_i)\, d\lambda_i dy}{\int \prod_t p\left(y_{it} \left| \lambda_i, \beta_0, \sigma_0^2, y_{i,t-1}\right.\right) f_0(\lambda_i)\, d\lambda_i} \right| d\Pi\left(\beta, \sigma^2, f \left| y_{1:N,0:T}\right.\right) d\beta d\sigma^2 df.
\end{aligned}$$

Note that along the same lines as part (i) "Random Effects: Result 1", the integrand

$$\begin{aligned}
&\left| \frac{\int \phi\left(y;\ \beta y_{iT} + \lambda_i, \sigma^2\right) \prod_t p\left(y_{it} \left| \lambda_i, \beta, \sigma^2, y_{i,t-1}\right.\right) f(\lambda_i)\, d\lambda_i dy}{\int \prod_t p\left(y_{it} \left| \lambda_i, \beta, \sigma^2, y_{i,t-1}\right.\right) f(\lambda_i)\, d\lambda_i} \right. \\
&\left. - \frac{\int \phi\left(y;\ \beta_0 y_{iT} + \lambda_i, \sigma_0^2\right) \prod_t p\left(y_{it} \left| \lambda_i, \beta_0, \sigma_0^2, y_{i,t-1}\right.\right) f_0(\lambda_i)\, d\lambda_i dy}{\int \prod_t p\left(y_{it} \left| \lambda_i, \beta_0, \sigma_0^2, y_{i,t-1}\right.\right) f_0(\lambda_i)\, d\lambda_i} \right| < \epsilon.
\end{aligned}$$

(iii) Correlated Random Effects: Result 1

As the posterior consistency for conditional density estimation is characterized by the joint

distribution over (λ_i, y_{i0}), the convergence of "joint" predictive distribution $(y_{i,T+1}, y_{i0})$ follows the same logic as part (i) "Random Effects: Result 1". Hence for any bounded continuous function $\tilde{\varphi}(y, y_{i0})$, and any $\epsilon > 0$, as $N \to \infty$,

$$\mathbb{P}\left(\begin{array}{l} f \in \mathcal{F},\ \left(\beta, \sigma^2\right) \in \mathbb{R} \times \mathbb{R}^+ : \\ \left| \int \tilde{\varphi}(y, y_{i0})\, f_{i,T+1}^{cond}\left(y|\beta, \sigma^2, f, y_{i,0:T}\right) q_0(y_{i0})\, dy_{i0} dy \right. \\ \left. - \int \tilde{\varphi}(y, y_{i0})\, f_{i,T+1}^{oracle}(y|y_{i0})\, q_0(y_{i0})\, dy_{i0} dy \right| < \epsilon \end{array} \middle| y_{1:N,0:T} \right) \to 1,\ a.s.$$

where

$$\begin{aligned} & \left| \int \tilde{\varphi}(y, y_{i0})\, f_{i,T+1}^{cond}\left(y|\beta, \sigma^2, f, y_{i,0:T}\right) q_0(y_{i0})\, dy_{i0} dy - \int \tilde{\varphi}(y, y_{i0})\, f_{i,T+1}^{oracle}(y|y_{i0})\, q_0(y_{i0})\, dy_{i0} dy \right| \\ = & \left| \frac{\int \tilde{\varphi}(y, y_{i0})\, \phi\left(y;\ \beta y_{iT} + \lambda_i, \sigma^2\right) \prod_t p\left(y_{it} \left| \lambda_i, \beta, \sigma^2, y_{i,t-1}\right.\right) f(\lambda_i|y_{i0})\, q_0(y_{i0})\, d\lambda_i dy_{i0} dy}{\int \prod_t p\left(y_{it}|\lambda_i, \beta, \sigma^2, y_{i,t-1}\right) f(\lambda_i|y_{i0})\, q_0(y_{i0})\, d\lambda_i dy_{i0}} \right. \\ & \left. - \frac{\int \tilde{\varphi}(y, y_{i0})\, \phi\left(y;\ \beta_0 y_{iT} + \lambda_i, \sigma_0^2\right) \prod_t p\left(y_{it} \left| \lambda_i, \beta_0, \sigma_0^2, y_{i,t-1}\right.\right) f_0(\lambda_i|y_{i0})\, q_0(y_{i0})\, d\lambda_i dy_{i0} dy}{\int \prod_t p\left(y_{it} \left| \lambda_i, \beta_0, \sigma_0^2, y_{i,t-1}\right.\right) f_0(\lambda_i|y_{i0})\, q_0(y_{i0})\, d\lambda_i dy_{i0}} \right| . \end{aligned} \tag{C.11}$$

However, it is more desirable to establish the convergence of "conditional" predictive distribution $y_{i,T+1}|y_{i0}$, i.e. for any bounded continuous function on y, $\varphi(y)$ and any $\epsilon > 0$, as $N \to \infty$,

$$\mathbb{P}\left(\begin{array}{l} f \in \mathcal{F},\ \left(\beta, \sigma^2\right) \in \mathbb{R} \times \mathbb{R}^+ : \\ \left| \int \varphi(y)\, f_{i,T+1}^{cond}\left(y|\beta, \sigma^2, f, y_{i,0:T}\right) dy - \int \varphi(y)\, f_{i,T+1}^{oracle}(y|y_{i0})\, dy \right| < \epsilon \end{array} \middle| y_{1:N,0:T} \right) \to 1,\ a.s.$$

where

$$\begin{aligned} & \left| \int \varphi(y)\, f_{i,T+1}^{cond}\left(y|\beta, \sigma^2, f, y_{i,0:T}\right) dy - \int \varphi(y)\, f_{i,T+1}^{oracle}(y|y_{i0})\, dy \right| \\ = & \left| \frac{\int \varphi(y)\, \phi\left(y;\ \beta y_{iT} + \lambda_i, \sigma^2\right) \prod_t p\left(y_{it} \left| \lambda_i, \beta, \sigma^2, y_{i,t-1}\right.\right) f(\lambda_i|y_{i0})\, d\lambda_i dy}{\int \prod_t p\left(y_{it}|\lambda_i, \beta, \sigma^2, y_{i,t-1}\right) f(\lambda_i|y_{i0})\, d\lambda_i} \right. \\ & \left. - \frac{\int \varphi(y)\, \phi\left(y;\ \beta_0 y_{iT} + \lambda_i, \sigma_0^2\right) \prod_t p\left(y_{it} \left| \lambda_i, \beta_0, \sigma_0^2, y_{i,t-1}\right.\right) f_0(\lambda_i|y_{i0})\, d\lambda_i dy}{\int \prod_t p\left(y_{it} \left| \lambda_i, \beta_0, \sigma_0^2, y_{i,t-1}\right.\right) f_0(\lambda_i|y_{i0})\, d\lambda_i} \right| . \end{aligned} \tag{C.12}$$

Set $\tilde{\varphi}(y, y_{i0}) = \frac{\varphi(y)}{q_0(y_{i0})}$. Note that $q_0(y_{i0})$ is continuous and bounded below due to condition 2-b in Proposition 4.16, so $\tilde{\varphi}(y, y_{i0})$ is a bounded continuous continuous function. Then, the right hand side of equation (C.11) coincides with the right hand side of equation (C.12), so we achieve the convergence of "conditional" predictive distribution $y_{i,T+1}|y_{i0}$.

(iv) Correlated Random Effects: Result 2

Combining (ii) and (iii) completes the proof. □

D Proofs for General Model

D.1 Identification

Proof. **(Proposition 5.6)**

Part (iii) follows Liu *et al.* (2016), which is based on the early work by Arellano and Bonhomme (2012). Part (ii) for cross-sectional heteroskedasticity is new.

(i) The identification of common parameters β is given by Assumption 5.5 (1).

(ii) Identify the distribution of shock sizes f^{σ^2}

First, let us perform orthogonal forward differencing, i.e. for $t = 1, \cdots, T - d_w$,

$$\begin{aligned}
\tilde{y}_{it} &= y_{it} - w_{i,t-1}' \left(\sum_{s=t+1}^{T} w_{i,s-1} w_{i,s-1}' \right)^{-1} \sum_{s=t+1}^{T} w_{i,s-1} y_{is}, \\
\tilde{x}_{i,t-1} &= x_{i,t-1} - w_{i,t-1}' \left(\sum_{s=t+1}^{T} w_{i,s-1} w_{i,s-1}' \right)^{-1} \sum_{s=t+1}^{T} w_{i,s-1} x_{i,s-1}.
\end{aligned}$$

Then, define

$$\begin{aligned}
\tilde{u}_{it} &= \tilde{y}_{it} - \beta' \tilde{x}_{i,t-1}, \\
\hat{\sigma}_i^2 &= \sum_{t=1}^{T-d_w} \tilde{u}_{it}^2 = \sigma_i^2 \chi_i^2.
\end{aligned}$$

where $\chi_i^2 \sim \chi^2 (T - d_w)$ follows an i.i.d. chi-squared distribution with $(T - d_w)$ degrees of freedom.

Note that Fourier transformation (i.e. characteristic functions) is not suitable for disentangling products of random variables, so I resort to the Mellin transform (Galambos and Simonelli, 2004). For a generic variable x, the Mellin transform of $f(x)$ is specified as

$$M_x(\xi) = \int x^{i\xi} f(x)\, dx,$$

which exists for all ξ.

Considering that $\sigma_i^2 | c$ and χ_i^2 are independent, we have

$$M_{\hat{\sigma}^2}(\xi | c) = M_{\sigma^2}(\xi | c)\, M_{\chi^2}(\xi).$$

Note that the non-vanishing characteristic function of σ^2 implies non-vanishing Mellin transform

$M_{\sigma^2}(\xi|c)$ (almost everywhere), so it is legitimate to take the logarithm of both sides,

$$\log M_{\hat{\sigma}^2}(\xi|c) = \log M_{\sigma^2}(\xi|c) + \log M_{\chi^2}(\xi).$$

Taking the second derivative with respect to ξ, we get

$$\frac{\partial^2}{\partial\xi\partial\xi'}\log M_{\sigma^2}(\xi|c) = \frac{\partial^2}{\partial\xi\partial\xi'}\log M_{\hat{\sigma}^2}(\xi|c) - \frac{\partial^2}{\partial\xi\partial\xi'}\log M_{\chi^2}(\xi).$$

The Mellin transform of chi-squared distribution $M_{\chi^2}(\xi)$ is a known functional form. In addition, we have

$$\begin{aligned}
\log M_{\sigma^2}(0|c) &= \log M_{\hat{\sigma}^2}(0|c) - \log M_{\chi^2}(0) = 0, \\
\frac{\partial}{\partial\xi}\log M_{\sigma^2}(0|c) &= \frac{\partial}{\partial\xi}\log M_{\hat{\sigma}^2}(0|c) - \frac{\partial}{\partial\xi}\log M_{\chi^2}(0) \\
&= i\left(\mathbb{E}\left(\log\hat{\sigma}^2\middle|\, c\right) - \mathbb{E}\left(\chi^2\middle|\, c\right)\right).
\end{aligned}$$

Based on Pav (2015),

$$\mathbb{E}\left(\chi^2\middle|\, c\right) = \log 2 + \psi\left(\frac{T-d_w}{2}\right),$$

where $\psi(\cdot)$ is the derivative of the log of the Gamma function.

Given $\log M_{\sigma^2}(0|c)$, $\frac{\partial}{\partial\xi}\log M_{\sigma^2}(0|c)$, and $\frac{\partial^2}{\partial\xi\partial\xi'}\log M_{\sigma^2}(\xi|c)$, we can fully recover $\log M_{\sigma^2}(\xi|c)$ and hence uniquely determine f^{σ^2}. Please refer to Theorem 1.19 in Galambos and Simonelli (2004) for the uniqueness.

(iii) Identify the distribution of individual effects f^{λ}

Define

$$\mathring{y}_{i,1:T} = y_{i,1:T} - \beta' x_{i,0:T-1} = \lambda_i' w_{i,0:T-1} + u_{i,1:T}.$$

Let $\mathring{Y} = \mathring{y}_{i,1:T}$, $W = w_{i,0:T-1}'$, $\Lambda = \lambda_i$ and $U = u_{i,1:T}$. The above expression can be simplified as

$$\mathring{Y} = W\Lambda + U.$$

Denote $F_{\mathring{Y}}$, F_{Λ} and F_U as the conditional characteristic functions for $\mathring{Y}$, Λ and U, respectively. Based on Assumption (5.5) (4), F_{Λ} and F_U are non-vanishing almost everywhere. Then, we obtain

$$\log F_{\Lambda}\left(W'\xi|c\right) = \log F_{\mathring{Y}}(\xi|c) - \log F_U(\xi|c).$$

Let $\zeta = W'\xi$ and $A_W = (W'W)^{-1}W'$, then the second derivative of $\log F_{\Lambda}(\zeta|c)$ is characterized by

$$\frac{\partial^2}{\partial\zeta\partial\zeta'}\log F_{\Lambda}(\zeta|c) = A_W\left(\frac{\partial^2}{\partial\xi\partial\xi'}\left(\log F_{\mathring{Y}}(\xi|c) - \log F_U(\xi|c)\right)\right)A_W'.$$

Moreover,

$$\begin{aligned} \log F_{\Lambda}(0|c) &= 0, \\ \frac{\partial}{\partial \zeta} \log F_{\Lambda}(0|c) &= i\mathbb{E}\left(A_W \mathring{Y} \middle| c \right), \end{aligned}$$

so we can pin down $\log \Lambda(\zeta|c)$ and f^{λ}. □

The proof of Proposition (5.8) for unbalanced panels follows in a similar manner.

D.2 Cross-sectional Heteroskedasticity

Proof. **(Proposition 5.9)**

(i) KL requirement

As λ and σ^2 are independent, we have

$$d_{KL}\left(f_0^{\lambda} f_0^{\sigma^2},\ f^{\lambda} f^{\sigma^2} \right) = d_{KL}\left(f_0^{\lambda},\ f^{\lambda} \right) + d_{KL}\left(f_0^{\sigma^2},\ f^{\sigma^2} \right).$$

Based on the observed sufficient statistics $\hat{\lambda} = \frac{1}{T}\sum_{t=1}^{T} y_{it}$ with corresponding errors $\hat{u} = \frac{1}{T}\sum_{t=1}^{T} u_{it}$, the KL requirement is: for all $\epsilon > 0$,

$$\Pi\left(\begin{array}{l} f \in \mathcal{F},\ f^{\sigma^2} \in \mathcal{F}^{\sigma^2} :: \\ \int f_0^{\lambda}\left(\hat{\lambda} - \hat{u}\right) \phi\left(\hat{u};\ 0, \frac{\sigma^2}{T}\right) f_0^{\sigma^2}\left(\sigma^2\right) \log \frac{\int f_0^{\lambda}\left(\hat{\lambda} - \hat{u}'\right) \phi\left(\hat{u};\ 0, \frac{\sigma^{2\prime}}{T}\right) f_0^{\sigma^2}\left(\sigma^{2\prime}\right) d\hat{u}' d\sigma^{2\prime}}{\int f^{\lambda}\left(\hat{\lambda} - \hat{u}'\right) \phi\left(\hat{u};\ 0, \frac{\sigma^{2\prime}}{T}\right) f^{\sigma^2}\left(\sigma^{2\prime}\right) d\hat{u}' d\sigma^{2\prime}} \\ \cdot\, d\hat{u} d\sigma^2 d\hat{\lambda} < \epsilon \end{array} \right) > 0.$$

As in the proof of Proposition 4.7 part (i), similar convexity reasoning can be applied to bound the KL divergence on y by $d_{KL}\left(f_0^{\lambda} f_0^{\sigma^2},\ f^{\lambda} f^{\sigma^2} \right)$. The sufficient conditions for KL properties on λ and l are listed in Lemmas 4.8 and E.1. Note that since the KL divergence is invariant under variable transformations, the KL property of the distribution of l is equivalent to the KL property of the distribution of σ^2.

(ii) Uniformly exponentially consistent tests

The alternative region can be split into the following two parts:

(ii-a) $f^{\sigma^2} \in U^{c}_{\epsilon', \Phi'}\left(f_0^{\sigma^2} \right)$

Orthogonal forward differencing yields $\tilde{y}_{it} \sim N\left(0, \sigma_i^2\right)$. Define $\hat{\sigma}_i^2 = \sum_{t=1}^{T-d_w} \tilde{y}_{it}^2 = \sigma_i^2 \chi_i^2$, where $\chi_i^2 \sim \chi^2(T - d_w)$ follows an i.i.d. chi-squared distribution with $(T - d_w)$ degrees of freedom. Here and below, I ignore the subscripts to simplify the notation.

Let $g^{\sigma^2}\left(\sigma^2\right) = f^{\sigma^2}\left(\sigma^2\right) - f_0^{\sigma^2}\left(\sigma^2\right)$. There are always tests if we observe σ^2, then for any g^{σ^2},

there exists a $\epsilon > 0$ such that

$$\int \left| g^{\sigma^2}\left(\sigma^2\right) \right| d\sigma^2 > \epsilon. \tag{D.1}$$

Similar to part (ii-b) in the proof of Proposition 4.7, here again I utilize the proof-by-contradiction technique. Suppose there is no test when $\hat{\sigma}^2$ is observed instead of σ^2, then there exist a $\tilde{g}^{\sigma}$ such that

$$\tilde{h}\left(\hat{\sigma}^2\right) = \int \tilde{g}^{\sigma^2}\left(\frac{\hat{\sigma}^2}{\chi^2}\right) f_{\chi^2}\left(\chi^2\right) d\chi^2 = 0 \text{ for all } \hat{\sigma}^2,$$

due to the continuity of $\tilde{h}$. Here I utilize the Mellin transform for products of random variables. As σ^2 and χ^2 are independent, we have

$$M_{\hat{\sigma}^2}(\xi) = M_{\sigma^2}(\xi) \cdot M_{\chi^2}(\xi) = 0 \text{ for all } \xi.$$

The Mellin transform of chi-squared distribution $M_{\chi^2}(\xi) \neq 0$, then

$$M_{\sigma^2}(\xi) = 0 \text{ for all } \xi.$$

Note that $M_{\sigma^2}(\xi)$ uniquely determines $\tilde{g}^{\sigma^2}\left(\sigma^2\right)$. Then, the inverse Mellin transform leads to

$$\tilde{g}^{\sigma^2}\left(\sigma^2\right) = 0 \text{ for all } \sigma^2,$$

which contradicts equation (D.1). Therefore, there are also tests distinguishing the true $f_0^{\sigma^2}$ from alternative f^{σ^2} even when we only observe $\hat{\sigma}^2$.

(ii-b') $f^{\sigma^2} = f_0^{\sigma^2},\ f^{\lambda} \in U^c_{\epsilon,\Phi}\left(f_0^{\lambda}\right)$

This is an intermediate step for part (ii-c). Once again I resort to proof by contradiction. Define $g^{\lambda}(\lambda) = f^{\lambda}(\lambda) - f_0^{\lambda}(\lambda)$. There are always tests if we observe λ, then for any g^{λ}, there exists a $\epsilon > 0$ such that

$$\int \left| g^{\lambda}(\lambda) \right| d\lambda > \epsilon. \tag{D.2}$$

Suppose there is no test when y is observed instead of λ, then there exist a $\tilde{g}^{\lambda}$ such that

$$\begin{aligned}
0 = \tilde{h}(y) &= \int \tilde{g}^{\lambda}(y-u)\,\phi\left(u;\ 0, \sigma^2\right) f_0^{\sigma^2}\left(\sigma^2\right) du d\sigma^2 \text{ for all } y \\
\Longrightarrow 0 = F_y(\xi) &= \int e^{-i\xi y} \tilde{g}^{\lambda}(y-u)\,\phi\left(u;\ 0, \sigma^2\right) f_0^{\sigma^2}\left(\sigma^2\right) du d\sigma^2 dy \\
&= \int e^{-i\xi(\lambda+\sigma v)} \tilde{g}^{\lambda}(\lambda)\,\phi\left(u;\ 0, \sigma^2\right) f_0^{\sigma^2}\left(\sigma^2\right) du d\sigma^2 d\lambda \\
&= F_{\lambda}(\xi) \cdot \int c_1 \exp\left(-c_2 \xi^2 \sigma^2\right) f_0^{\sigma^2}\left(\sigma^2\right) d\sigma^2 = 0 \text{ for all } \xi \\
\Longrightarrow F_{\lambda}(\xi) &= 0 \text{ for all } \xi \\
\Longrightarrow \tilde{g}^{\lambda}(\lambda) &= 0 \text{ for all } \lambda,
\end{aligned}$$

which contradicts equation (D.2). Therefore, there are also tests if we know $f_0^{\sigma^2}$ but only observe y.

(ii-b) $f^{\sigma^2} \in U_{\epsilon',\Phi'}\left(f_0^{\sigma^2}\right),\ f^{\lambda} \in U^c_{\epsilon,\Phi}\left(f_0^{\lambda}\right)$

Without loss of generality, let $\Phi = \{\varphi\}$ and $\varphi^\star$ be the corresponding test function when $f_0^{\sigma^2}$ is known as in case (ii-b'). Then, the difference between $\mathbb{E}_f(\varphi^\star)$ and $\mathbb{E}_{f_0}(\varphi^\star)$ is

$$\begin{aligned}
&\int \varphi^\star\left(\hat{\lambda}\right) f^{\lambda}\left(\hat{\lambda}-\hat{u}\right)\phi\left(\hat{u};\,0,\frac{\sigma^2}{T}\right) f^{\sigma^2}\left(\sigma^2\right) d\hat{u}d\sigma^2 d\hat{\lambda} - \int \varphi^\star\left(\hat{\lambda}\right) f_0^{\lambda}\left(\hat{\lambda}-\hat{u}\right)\phi\left(\hat{u};\,0,\frac{\sigma^2}{T}\right) f_0^{\sigma^2}\left(\sigma^2\right) d\hat{u}d\sigma^2 d\hat{\lambda} \\
&> \int \varphi^\star\left(\hat{\lambda}\right)\left(f^{\lambda}\left(\hat{\lambda}-\hat{u}\right) - f_0^{\lambda}\left(\hat{\lambda}-\hat{u}\right)\right)\phi\left(\hat{u};\,0,\frac{\sigma^2}{T}\right) f_0^{\sigma^2}\left(\sigma^2\right) d\hat{u}d\sigma^2 d\hat{\lambda} \\
&\quad - \left|\int \varphi^\star\left(\hat{\lambda}\right) f^{\lambda}\left(\hat{\lambda}-\hat{u}\right)\phi\left(\hat{u};\,0,\frac{\sigma^2}{T}\right)\left(f^{\sigma^2}\left(\sigma^2\right) - f_0^{\sigma^2}\left(\sigma^2\right)\right) d\hat{u}d\sigma^2 d\hat{\lambda}\right|.
\end{aligned}$$

Case (ii-b') implies that the first term is greater than some $\epsilon > 0$. Meanwhile, we can choose $\epsilon' = \epsilon/2$ and $\Phi' = \left\{\varphi'\left(\sigma^2\right) = 1\right\}$ for $U_{\epsilon',\Phi'}\left(f_0^{\sigma^2}\right)$ so that the second term is bounded by $\epsilon/2$. Hence, $\mathbb{E}_f(\varphi^\star) - \mathbb{E}_{f_0}(\varphi^\star) > \epsilon/2$, and $\varphi^\star$ is the test function with respect to the alternative sub-region $\left\{f^{\sigma^2} \in U_{\epsilon',\Phi'}\left(f_0^{\sigma^2}\right),\ f^{\lambda} \in U^c_{\epsilon,\Phi}\left(f_0^{\lambda}\right)\right\}$. □

E Extension: Heavy Tails

Lemma E.1 gives one set of conditions accommodating f_0^z with heavy tails using the Gaussian-mixture DPM prior. It follows Tokdar (2006) Theorem 3.3. The notation is slightly different from Tokdar (2006). Here G_0^z is defined on $\left(\mu_i^z, (\omega_i^z)^2\right)$, the mean and the variance, while Tokdar (2006) has the mean and the standard deviation as the arguments for G_0^z.

Lemma E.1. *(Tokdar, 2006)*

If f_0^z and the DP base distribution G_0^z satisfy the following conditions:

1. $\left|\int f_0^z(z)\log f_0^z(z)\,dz\right| < \infty$.
2. *For some* $\eta \in (0,1)$, $\int |z|^\eta f_0^z(z)\,dz < \infty$.
3. *There exist* $\omega_0 > 0$, $0 < b_1 < \eta$, $b_2 > b_1$, *and* $c_1, c_2 > 0$ *such that for large* $\mu > 0$,

$$\max\left\{\begin{array}{c} G_0^z\left(\left[\mu-\omega_0\mu^{\frac{\eta}{2}},\infty\right)\times\left[\omega_0^2,\infty\right)\right),\ G_0^z\left([0,\infty)\times\left(\mu^{2-\eta},\infty\right)\right), \\ G_0^z\left(\left(-\infty,-\mu+\omega_0\mu^{\frac{\eta}{2}}\right]\times\left[\omega_0^2,\infty\right)\right),\ G_0^z\left((-\infty,0]\times\left(\mu^{2-\eta},\infty\right)\right) \end{array}\right\} \geq c_1\mu^{-b_1},$$

$$\max\left\{\begin{array}{c} G_0^z\left((-\infty,\mu)\times\left(0,\exp\left(2\mu^\eta-1\right)\right)\right), \\ G_0^z\left((-\mu,\infty)\times\left(0,\exp\left(2\mu^\eta-1\right)\right)\right) \end{array}\right\} > 1-c_2\mu^{-b_2}.$$

Then, $f_0^z \in KL(\Pi^z)$.

The next lemma extends Lemma E.1 to the multivariate case. Then, Proposition E.3 largely parallels Proposition (5.10) with different condition sets for the KL property, which accounts for heavy tails in the true unknown distributions..

Lemma E.2. *(Heavy Tails: Multivariate)*

If f_0^z and the DP base distribution G_0^z satisfy the following conditions:

1. $\left|\int f_0^z(z)\log f_0^z(z)\,dz\right| < \infty$.
2. *For some* $\eta \in (0,1)$, $\int \|z\|^\eta f_0^z(z)\,dz < \infty$.
3. *There exist* $\omega_0 > 0$, $0 < b_1 < \eta$, $b_2 > b_1$, *and* $c_1, c_2 > 0$ *such that for large* $\mu > 0$, *for all directional vectors* $\|z^*\| = 1$,

$$\max\left\{\begin{array}{c} G_0^z\left(\left[\mu-\omega_0\mu^{\frac{\eta}{2}},\infty\right)\times\left[\omega_0^2,\infty\right)|z^*\right),\ G_0^z\left([0,\infty)\times\left(\mu^{2-\eta},\infty\right)|z^*\right), \\ G_0^z\left(\left(-\infty,-\mu+\omega_0\mu^{\frac{\eta}{2}}\right]\times\left[\omega_0^2,\infty\right)|z^*\right),\ G_0^z\left((-\infty,0]\times\left(\mu^{2-\eta},\infty\right)|z^*\right) \end{array}\right\} \geq c_1\mu^{-b_1},$$

$$\max\left\{\begin{array}{c} G_0^z\left((-\infty,\mu)\times\left(0,\exp\left(2\mu^\eta-1\right)\right)|z^*\right), \\ G_0^z\left((-\mu,\infty)\times\left(0,\exp\left(2\mu^\eta-1\right)\right)|z^*\right) \end{array}\right\} > 1-c_2\mu^{-b_2},$$

where $G_0^z(\cdot|z^*)$ represents the conditional distribution that is induced from $G_0^z(\cdot)$ conditional on the direction z^*.

Then, $f_0^z \in KL(\Pi^z)$

Proposition E.3. *(General Model: Random Coefficients II)*

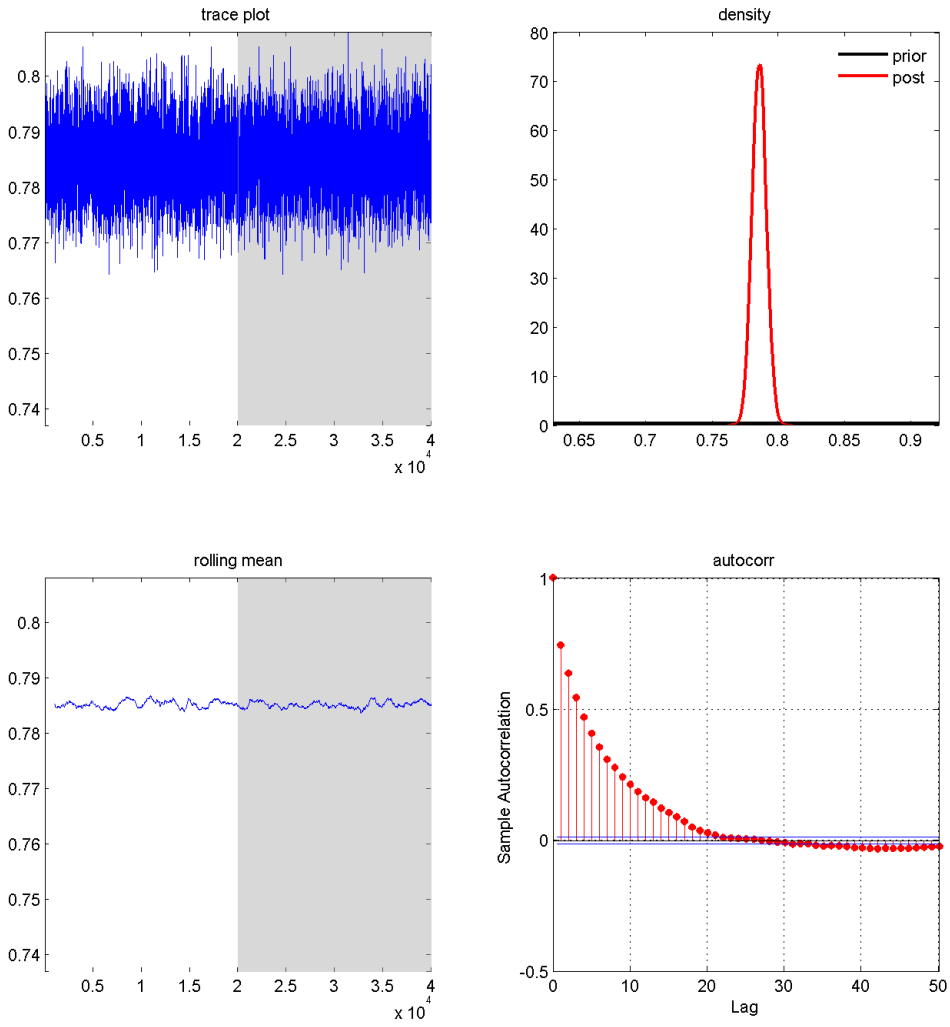
Suppose we have:

1. *Assumptions 5.3, 5.5 (3-4), 5.7, and 4.10.*
2. *Lemma E.2 on λ and Lemma E.1 on l.*
3. $\beta_0 \in \operatorname{supp}\left(\Pi^\beta\right)$.

Then, the posterior is weakly consistent at $\left(\beta_0, f_0^\lambda, f_0^{\sigma^2}\right)$.

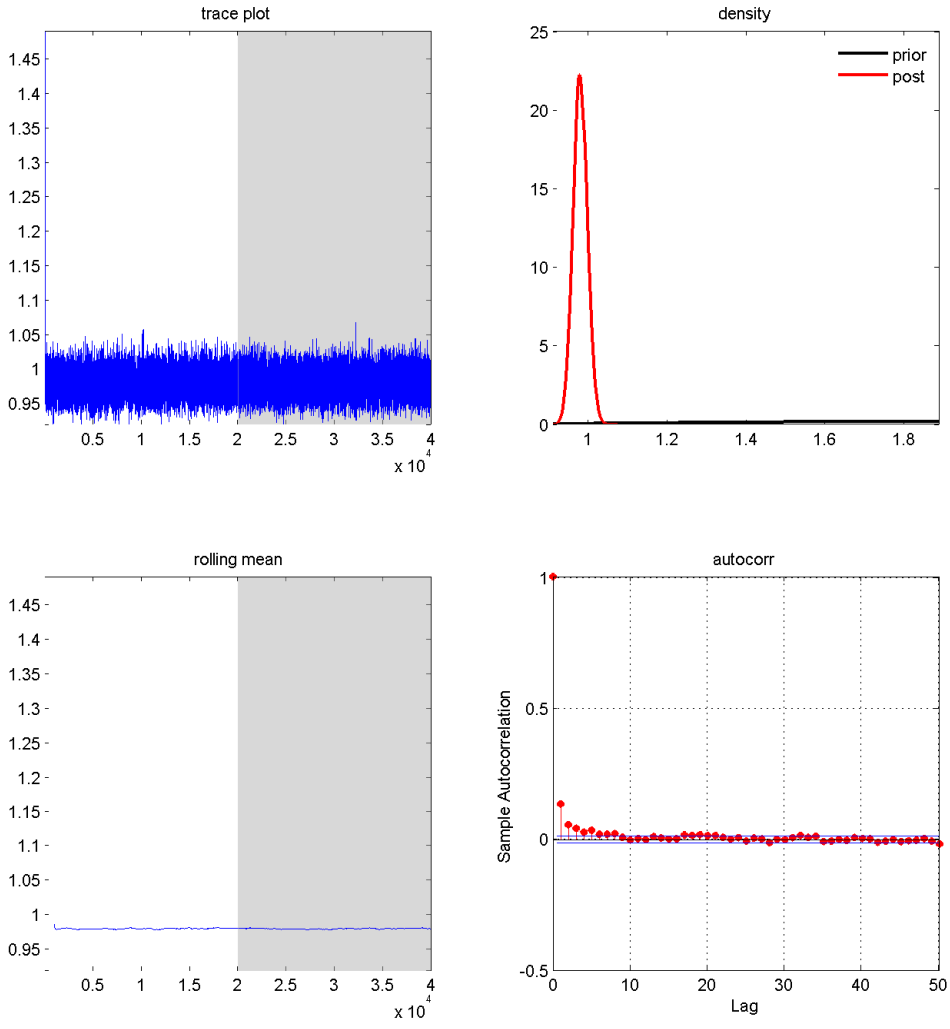
F Simulations

Figure F.1: Convergence Diagnostics: β



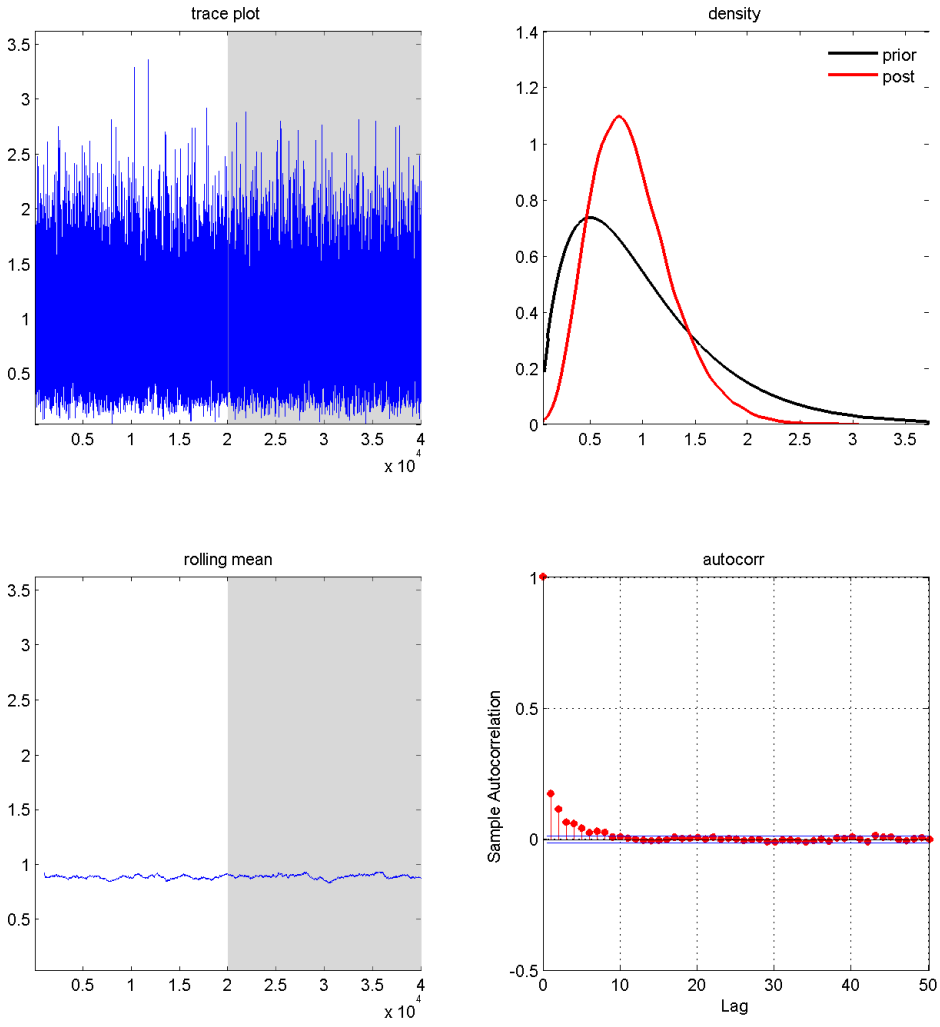
For each iteration s, rolling mean is calculated over the most recent 1000 draws.

Figure F.2: Convergence Diagnostics: σ^2



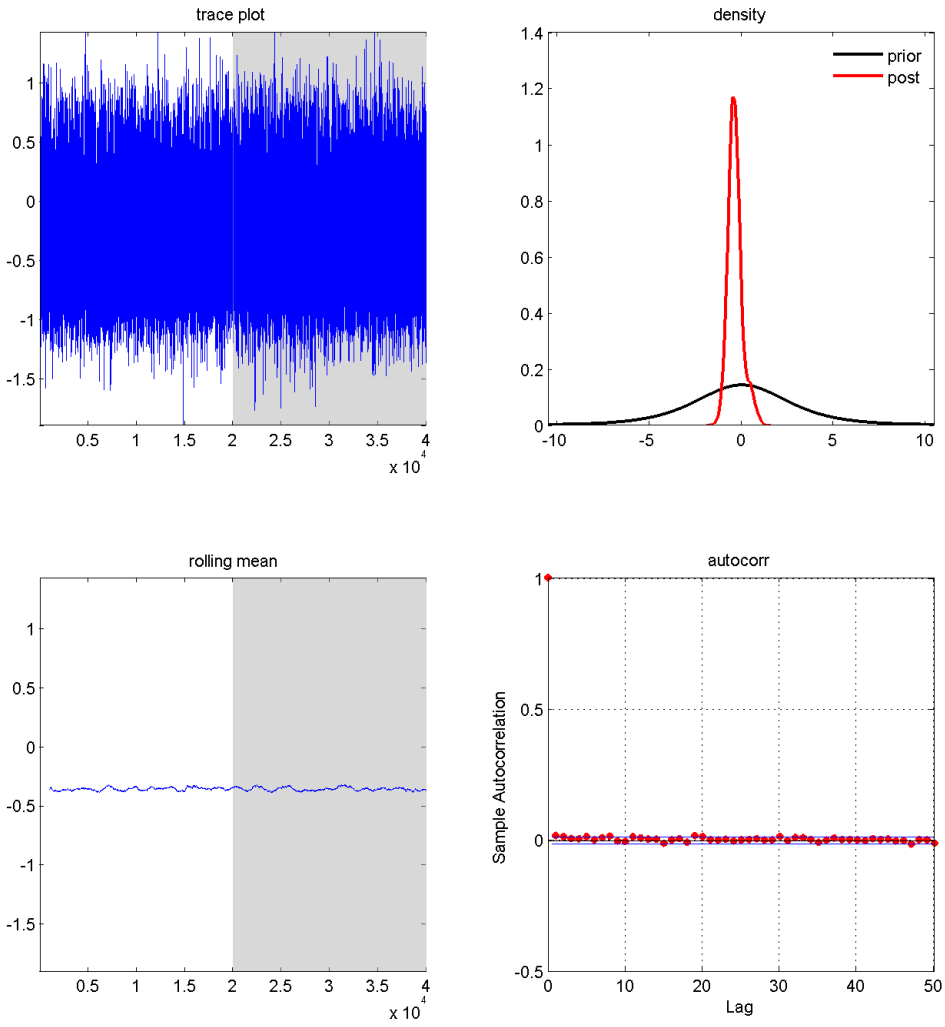
For each iteration s, rolling mean is calculated over the most recent 1000 draws.

Figure F.3: Convergence Diagnostics: α



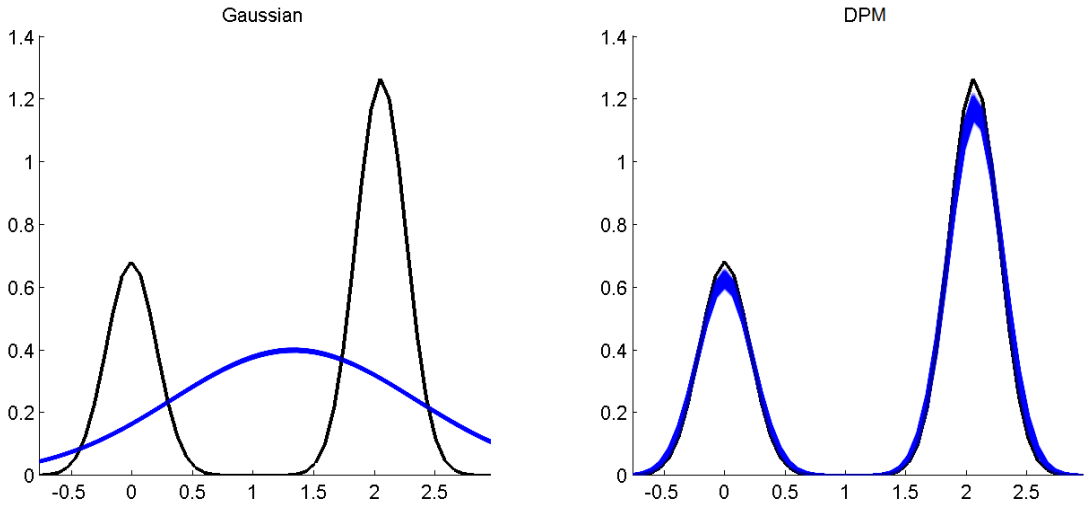
For each iteration s, rolling mean is calculated over the most recent 1000 draws.

Figure F.4: Convergence Diagnostics: λ_1



For each iteration s, rolling mean is calculated over the most recent 1000 draws.

Figure F.5: f_0 vs $\Pi\left(f \mid y_{1:N,0:T}\right)$: Baseline Model, $N = 10^5$



The black solid line represents the true λ_i distribution, f_0. The blue bands show the posterior distribution of f, $\Pi\left(f \mid y_{1:N,0:T}\right)$.